University of California

Los Angeles

The learnability of tones from the speech signal

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Linguistics

by

Kristine Mak Yu

2011

© Copyright by

Kristine Mak Yu

2011

The dissertation of Kristine Mak Yu is approved.

Abeer Alwan

Bruce Hayes

Sun-Ah Jun

Patricia Keating

Kie Ross Zuraw

Edward Stabler, Committee Co-chair

Megha Sundara, Committee Co-chair

University of California, Los Angeles

2011

To my family and teachers

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am indebted to many people for the completion of this thesis. I am especially grateful to my co-advisers, Edward Stabler and Megha Sundara. To Ed, I am grateful for guiding my vision of this project and of my scientific goals in general. I am grateful for his willingness to think about questions spanning such a wide breadth so clearly and deeply with me, and for putting up with all my questions and notes. To Megha, I owe my initial and continuing interest in infant language research which largely influenced how I framed my research questions in the thesis. I am particularly grateful for her invaluable guidance in experimental design and analysis and for her efforts to keep me and my research and writing focused.

I have been so fortunate to benefit from the guidance of the rest of my committee members as well. Bruce Hayes first suggested that I take lexical tones as a starting point for studying prosodic acquisition, was always willing to lend an ear whenever I wanted to talk, and asks stimulating questions that still keep me thinking. Sun-Ah Jun taught me about prosody and also inspired a whole large-scale project on melodic rhythm and prosodic acquisition which didn't make it into the thesis. Patricia Keating taught me about voice quality, which came up as an issue that never went away as soon as I started data collection, and she guided me in tackling it and all sorts of other phonetic issues. I am also grateful to her for providing funding for a number of my experiments, through NSF grant BCS-0720304 to her, Abeer Alwan, and Jody Kreiman. Kie Zuraw has been a joy to discuss anything with, from nitty gritty phonetic detail to mathematical models to treadmill desks. Abeer Alwan and her students Gang Chen, Wei Chu, and especially Yen-Liang Shue provided invaluable technical support on speech processing, especially on voice quality analysis and pitch detection. Jody Kreiman helped me think about voice quality and the design of the voice quality experiments in Cantonese.

I am grateful for all my other teachers at UCLA, as well, and classmates and staff, who make the department community so warm and intellectually lively. I am grateful to all the members of the UCLA phonetics lab, some now alumni, especially Roy Becker, Jason Bishop, Marc Garellek, Sameer Khan, Jianjing Kuang, Grace Kuo, Molly Shilman, and Chad Vicenik for all their help over the years (and for sharing lodgings for conferences), and my cohort members Byron Ahn, Ben George, Robyn Orfitelli, and J'aime Roemer for all their support. Our irreplaceable departmental staff, lab managers, and engineer Anya Essiounina, Lisa Harrington, Melanie Levin, Kristi Hendrickson, Adrienne Scutellaro, and Henry Tehrani kept everything running smoothly. Outside of the UCLA community, I am also grateful to Keelan Evanini, John Kingston, Mark Liberman, Morgan Sonderegger, and Colin Wilson for discussion, suggestions, and technical help, as well as audiences at the UCLA phonetics and phonology seminars, ASA in Spring 2010 and 2011, the Berkeley tone workshop, ICPhS 2011, LSA in January 2011, NELS 41, and UMass Amherst.

For getting me into linguistics, I am grateful to my very first linguistics teachers, Barb Kelly and Will Leben and Liz Coppock and Andrew Koontz-Garboden at Stanford, and Will in particular for mentoring me through my transition into linguistics. I am also so thankful to my mentors in chemistry and biology: Peter Oefner, Grace Rosenquist, John Ross, and Antonio Sanchez-Torralba, who taught me how to do scientific research and fully supported me when I switched fields.

It would not have been possible to collect and analyze the data for this thesis without help from many, many collaborators. For the Bole data, I am grateful to Russ Schuh and Gimba, who designed and recorded the stimuli in Nigeria in Winter 2009, and to Russ Schuh for funding the recording with NSF grant BCS-0553222. I am also grateful to Russ for introducing me to Balkan music—a happy diversion—and all our chats about tone and running and music.

For the Cantonese work, I had help from a wonderful group of research assistants and collaborators. Cindy Chan, Hiu Wai (Irene) Lam, Shing Yin (Celine) Li, and Cedric Loke helped me design, record, and analyze the production data in Summer and Fall 2009. The first experiment on creaky voice in Cantonese was Irene's honors thesis, and the second, as well as the other two Cantonese perception experiments, would have been impossible without Irene's tireless efforts to help me pilot and run the experiments all by herself in Hong Kong in Summer 2010—I am also grateful to Eric Zee for letting us occupy his lab's soundbooth at the City University of Hong Kong for a whole month! Prairie Lam stepped in in the nick of time to help finish recruiting and running subjects for perception experiments in Fall 2010 and also helped with production data analysis, and Antonio Sou helped with piloting. Vincie Ho also graciously helped check the accuracy of Cantonese tonal productions when she was a visiting scholar in the department.

For my interest in Hmong, I am thankful to Pat Keating and Christina Esposito. Sherrie Yang helped teach me what little Hmong I know and helped coordinate my contacts with the Los Angeles Hmong community; Chou Khang was my patient consultant for stimuli design, and Phong Yang and Sean Fulop of CSU Fresno hosted me in Fresno in Spring 2010—Phong helped me with finding subjects even though his wife went into labor the day I arrived in Fresno. I am also grateful to Porchoua Her for coordinating my recordings at the Center for New Americans and to the center for hosting me.

Grace Kuo helped me with designing the Mandarin stimuli, and Jianjing

Kuang helped me record the Beijing Mandarin in Beijing in Summer 2009. Samantha Chan, Calvin Cheng, Chris Fung, J'aime Roemer, Antonio Sou, and Joanna Wang helped with the data analysis.

There are two languages which did not make it into the thesis since work on them is still in progress—Igbo and Yoruba—and I would like to thank Echezona Okoli and Moses Chikwe for language consultation for Igbo, and Moses Adegbola, Sunday Komolafe, and Saheed Raheem for all their help with Yoruba, and also Douglas Pulleyblank and Akin Akinlabi for answering my questions about Yoruba. I am also grateful to my dedicated undergraduate research assistants Afton Coombs, J. Francisco DeRuvira, Elyssa Dudley, Wai Lam Fu, Gina Ger, Emily Ho, Katie Redderson-Lear, Amanda Ritchart, Grace Tsai, and Brent Woo for work on other African tone languages and intonational languages and tonal typology.

I am of course also so appreciative of all the native speakers who worked with me and participated in experiments, some of whom wish to remain anonymous, and: Aishatu, Baba Ali, Gimba, Jummai, Madu; Cedric Loke, Kin, Ho Ming Tam, Tek Ngai Wong; Beu Her, Chue Her, Porchoua Her, Nalee Thao, Mai Summer Vue, Jeff Xiong, Yeng Xiong, Cyndee Lor-Yang, Bee Yang, Phong Yang, Yer Yang; Ben-Yue Chang, Che-kang Chang, Li Wei Chen, Po-Wei Chen, Meng-Ning Cheng, Nelson Cheng, Jih-Chung Fan, Grace Kuo, Jao Yu Lee, Yu-Lun Lin, Shih Peng Sun, Liliane Hsin-Yi Teng, Pao-Yi Tseng.

There are many funding sources which made this work possible: in addition to NSF grant BCS-0720304 to Patricia Keating, Abeer Alwan, and Jody Kreiman and NSF grant BCS-0553222 to Russ Schuh, I am also grateful for a NSF graduate fellowship, a Ladefoged Scholarship from the UCLA Linguistics Department for the Hmong fieldwork, and funding for subjects and consultants from the linguistics department. For general funding during my time as a graduate student, I am also grateful for funding from the Graduate Research Mentorship Program and Summer Graduate Research Mentorship Program sponsored by the UCLA graduate division and research assistantships from Pat Keating and Kie Zuraw.

Outside of academics, I am grateful to my family members for all their support and encouragement, especially my sister Terri who helped me over the years in so many ways. I am thankful for my roommate Kathy Kornei who kept me sane and shared many delicious cooking and market adventures with me; all my meditation, swimming and yoga instructors, including Marvin Belzer, Brenda Johnson, Drew Porter, Andrea Wagner, and Diana Winston, and my running buddies Jason Bishop, Grace Kuo, and Chacha Mwita; the Varimezov family and my bandmates and the singers in the UCLA Balkan ensemble, and Katherine Hill for the magic prospectus bowl and the dissertation dragon formerly known as C. Phil Strange.

Finally, I hope that anyone I have regrettably forgotten to include will understand that I am grateful to them as well. In fact, since my thesis would not have been possible without the efforts of all the scientists who developed ideas, software, mathematical models, grammars, dictionaries, etc., that I have relied on and without the help of so many people who have nourished and taught me throughout my life, it is impossible to list everyone I am thankful to in any case. To all of you, I thank you.

Vita

1984	Born, San Jose, California, USA.
2001–2004	B.S., Chemistry Stanford University Stanford, California
2006–2008	M.A., Linguistics University of California, Los Angeles Los Angeles, California

Abstract of the Dissertation

The learnability of tones from the speech signal

by

Kristine Mak Yu

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2011

Professor Megha Sundara, Co-chair

Professor Edward Stabler, Co-chair

It is an unremarkable matter of course but a remarkable miracle of human cognition that children learning tonal languages learn maps from the speech signal to the abstract phonological tone concepts of their native language, which could be any tone language of the world. As an initial step for understanding how children learn tonal maps, this thesis focuses on working toward a characterization of what it is that is being learned—the class of possible maps from the speech signal to tonal categories in natural language. By studying the structure of this class of tonal maps, we can assess the learnability of the class under a mathematically precise criterion for successful feasible learning. Characterizing the learning problem as feasibly learnable is a fruitful direction for elucidating the human learning problem.

Since the structure of tonal maps is conditioned on the phonetic space in which they are defined, the focus of this thesis is determining an appropriate phonetic parameterization of the speech signal for the domain of the tonal maps and for representation of the data to the learner. We do this by assessing the separability of tonal categories in different phonetic spaces. Studying the structure of the class of possible tonal maps necessitates studying tonal maps in a range of languages, so we study tonal maps using a sample of cross-linguistic tonal production data we collected in Bole, Beijing Mandarin, Cantonese, and White Hmong and with a series of perception experiments we performed in Cantonese.

The bulk of the thesis motivates the inclusion of particular information from the speech signal, since the phonetic realization of linguistic tone is widely believed to be limited to a single dimension of fundamental frequency, the acoustic correlate of pitch. We show evidence from human perceptual experiments and computational modeling: (i) motivating a temporal domain from the speech signal for tonal maps beyond the span of a single syllable, and (ii) demonstrating that voice source parameters beyond f0 must be included for characterizing phonetic spaces for tonal maps in a wide range of languages.

While these results indicate potential sources of complexity for tonal maps, we also show that coarse temporal resolution in sampling of the relevant parameters from the speech signal suffices for good tonal category separability, hinting at potential structure in tonal maps. Human listeners identify tones degraded to be coarsely sampled at a comparable level of accuracy to that for intact tones in Cantonese, and classification by machine with acoustic parameter spaces defined only over a few real values shows a near partition of the phonetic space in the sample of languages studied. The potential structure in tonal maps suggested by these results is consistent with feasible learnability of tonal maps.

CHAPTER 1

Introduction

In the days when the Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

What are you doing? asked Minsky.
I am training a randomly wired neural net to play Tic-tac-toe, Sussman replied.
Why is the net wired randomly? asked Minsky.
I do not want it to have any preconceptions of how to play, Sussman said.

Minsky then shut his eyes.
Why do you close your eyes? Sussman asked his teacher.
So that the room will be empty.
At that moment, Sussman was enlightened.
(Appendix to the Jargon File, (on-line hacker Jargon File, 2003))

. . . an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied. (Marr, 1982, 27)

Overview

It is an unremarkable matter of course but a remarkable miracle of human cognition that children learning tonal languages learn maps from the speech signal to the abstract phonological tone concepts of their native language, which could be any tonal language of the world.

A natural question is: *how* do children learn tonal maps—what is the computational procedure underlying this learning process? This is a question about the *algorithmic* implementation of the learner. But an algorithm is a procedure that transforms input into output, given a well-specified set of inputs, set of outputs, and relation between the two (Cormen et al., 2001, 5): the characterization of a learning algorithm for maps from the speech signal to tonal concepts is predicated on the characterization of the input-output map—the learning function.

Unfortunately, at this point in time, we cannot precisely characterize any component of the learning function for tonal maps. We may "know" that children learn tonal maps, but we do not know what kind of evidence is used in learning tonal maps—the domain of the function; we do not know what speakers know about tonal maps—the range of the function; we do not know how the evidence and tonal maps are related—the map between the input and output values; we do not even know in exactly what sense it means to say that the tonal maps are "learned".

Thus, as an initial step for understanding how children learning tonal languages learn maps from the speech signal to tonal concepts, this thesis focuses on characterizing *what* is being computed by the learner. We seek to elucidate the range of the learning function for maps from the speech signal to tonal concepts—the hypothesis space of the learner—and in particular, potential *structure* in the hypothesis space. Such structure directs our study of what a learning mechanism for tonal maps could be; it gives us clues for how the learner might generalize from the finite number of instances heard in the input to the infinitude of elements defined by tonal concepts.

The strategy we take is to study aspects of tonal maps in a sample of speakers of a sample of tonal languages. From the maps we study, we can better understand what the space of the class of possible tonal maps for human languages is. An understanding of the structure in the hypothesis space is necessary and sufficient to characterize if the class of tonal maps is feasibly learnable, where the notion of successful learning is mathematically precisely defined (Blumer et al., 1989; Valiant, 1984; Vapnik and Chervonenkis, 1971). If the class does not appear to be feasibly learnable, then we would reconsider the proposed structural restrictions on possible tonal maps. Even if the proposed criterion for learnability might not be exactly the relevant one for human learning so that our work does not bear in the most direct way on the human learning problem, we expect work towards characterizing the learning problem as feasibly learnable to provide necessary direction for elucidating the human learning problem.

The body of this thesis addresses what information from the speech signal is referenced in tonal maps. This tells us about the domain of maps from (a subset of) information from the speech signal to tonal concepts. The bulk of our work motivates the *inclusion* of particular information from the speech signal in the domain for tonal maps (Ch. 2, 3, 5), but Ch. 4 and 5 also provide a glimpse of structure in the class of tonal maps that would limit its complexity. We emphasize the inclusion of information from the speech signal in the thesis because the phonetic realization of linguistic tone is widely believed to be simple in the sense that it is limited to a single dimension of fundamental frequency, f0 (the physical correlate of the auditory percept of pitch) (Gauthier et al., 2007, 82),(Hyman, 2010, 1).

But even an entirely f0-based parameterization of tone can be highly multidimensional, since we may choose to sample f0-based values arbitrarily densely in time (Ch. 4) from an arbitrarily large temporal window of the speech signal (Ch. 2), and we may choose multiple ways to parametrize f0, e.g. with f0 height values and with f0 velocity values (Ch. 5). We show that human listeners and machines benefit from local contextual information from both neighboring and preceding syllables in tonal identification, motivating a temporal domain from the speech signal for tonal maps beyond the span of a single syllable (Ch. 2). Furthermore, we demonstrate that f0-based parameters are insufficient for characterizing tones in a wide range of tonal languages because: (i) there are tone languages in which f0-based parameters alone are insufficient for contrasting tones, and even in tone languages where f0-based parameters are thought to be sufficient for distinguishing all tonal contrasts, (ii) f0 interacts with other voice source parameters and cannot be extracted from the speech signal independent of these other parameters (Chs. 4, 2, 3, 5) and (iii) listeners are sensitive to other voice source parameters (Ch. 3).

However, there is structure in the class of possible tonal maps: we show that coarse temporal resolution in sampling of the relevant parameters suffices for good tonal concept separability—human listeners identify tones degraded to be coarsely sampled at a comparable level of accuracy to that for intact tones (Ch. 4), and classification by machine with acoustic parameter spaces defined only over a few real values shows a near partition of the phonetic space in a range of languages (Chs. 4, 5).

In this situating chapter, we introduce terminology used in the thesis (§1.1.1), describe the learning problem in the context of phonological category acquisition, motivate and define the study of the target of learning, the map from the phonetic space to phonological categories, (§1.1.2), and describe the larger research questions (§1.1.3) and methodological abstractions taken in the thesis (§1.1.4) in §1.1. Then, we explicate our particular model system (§1.2) for studying tonal maps. We conclude by discussing what aspects of the parameterization of the speech signal for phonological maps our model system allows us to study in comparison to previous model systems for the acquisition of phonological categories and also specify which particular aspects we focus on in the thesis (§1.3).

1.1 Preliminaries

To begin with, we briefly introduce some terminology that we will use throughout the thesis.

1.1.1 Some terminology

1.1.1.1 Learners and learnability

A **learner** is a **map** (a function) from a collection of possible **data** sets (the function's domain) to a class of target **concepts** (the function's range). The data sets are given as sets of **examples**, instances of each target concept, and the concepts to be learned are **categories** (§1.1.1.2). We also call the learner's range, the class of target concepts, the **hypothesis space** of the learner.

For this thesis, the target concepts to be learned are what we call **phonolog-**

ical maps—specifically, **tonal maps**—and we also refer to the hypothesis space as the **class of possible (tonal) maps**. Explaining what phonological maps are is the topic of the following section §1.1.2. Briefly, they are maps (functions) from parameter vectors defining the space from which the examples for the learner are drawn to phonological categories (§1.1.1.2).

The definition of **learnability** depends on the definition of the criteria for successful learning. Learnability is a property of a hypothesis space rather than a specific learner (learning function), since a hypothesis space does not uniquely pick out a learner. Some criteria for successful learning require strictly that the learner converge *exactly* on the target of learning for any target in the hypothesis space. For modeling human learning, this may be too strict a criterion, so the convergence can be relaxed to allow some deviation from the target in some mathematically precise way. In addition, criteria for successful learning that allow the learner sets of examples that are infinite in size do not consider resource limitations that may be applicable in human learning. When we discuss whether something is **feasibly learnable**, we informally refer to a criteria for successful learning that takes into account limitations in resources for the learner that are potentially relevant for human learning.

1.1.1.2 Parameters, parameter spaces, and categories

When we refer to a **parameter** in this thesis, we mean a real-valued parameter. A set of parameters comprise a **parameter space** in $\mathbb{R}^d$, where d is the **dimensionality** of the parameter space and is simply the cardinality of the parameter set, the number of parameters in the set. We call the specific kind of parameter space discussed in the thesis a **phonetic parameter space** because the parameters are phonetic in nature. By **phonetic**, we refer to any property involved in the process of the transmission of speech sounds between speakers and listeners—a phonetic parameter may be some quantity that is measurable from the acoustic or articulatory record of the physical speech signal, e.g. the first formant at the midpoint of a vowel or the maximum displacement of the tongue tip during an obstruent, or some quantitative property of audition, e.g. pitch, or a yet higher order measurable property of linguistic cognition.

In the thesis, we refer to maps from phonetic parameter spaces to phonological **concepts** or **categories**. We use the terms **concept** and **category** interchangeably. When we refer to either, we do not mean the label for a category, such as "Tone 1" or "the vowel /a/"; we are actually referring to a mapping from the parameter space to a discrete element, a region (a set of points) over the parameter space labeled with the category name. "A category is a mental construct which relates two levels of representation, a discrete level and a parametric level" (Pierrehumbert, 2003b, 119).

When we use the term **tone** we mean *lexical* tone unless otherwise specified, and in particular, we intend to refer to a lexical tonal *phonological category*. In this thesis, we consider only maps to lexically contrastive phonological categories, i.e. *phonemes*.

1.1.1.3 Category shapes and distributions

The consideration of the shape and distribution of concepts is crucial for assessing the learnability of the class of concepts. One important property of a concept, if it is a bounded region in a parameter space, is if it is **connected**. This means that there are no "holes" in the region—all the denumerably infinitely many points in the parameter space enclosed in the region are labeled as instances of the concept. A subset of connected sets in the parameter space (regions) are also

convex sets. A set of points if convex if a line segment may be drawn between any two points in the set without leaving the region bounded by the set.

For the thesis, an important property of a phonetic space is how **separable** the tonal concepts are in the space. Here, the geometric intuitive understanding of separability is accurate: if two concepts overlap in the phonetic space, they are not well-separated; if they are far apart (in Euclidean distance), they are well-separated. Two concepts are **linearly separable** in a space if one can separate them with a line, if the space is in $\mathbb{R}^2$, or more generally, with a hyperplane, the higher-dimensional analogue of a line, in spaces in $\mathbb{R}^d$ for $d > 2$. We discuss linear separability further in Chapters 2, 4, and 5.

1.1.1.4 Samples and sampling resolution

The realization of a tone as a physical speech signal event unfolds in time. Therefore, the phonetic space for tones might very well include a sequence of parameter vectors $\langle \vec{v_1}, \vec{v_2}, \ldots, \vec{v_i}, \ldots, \vec{v_n} \rangle$ extracted over some finite number of timesteps t_i for $\{i \in \mathbb{N}^+ | 1 \leq i \leq n\}$. We call a parameter vector extracted at some timestep t_i a **sample**. The calculation of a parameter vector $\vec{v_i}$ may occur over some finitely bounded temporal window, **a frame**, rather than at an instantaneous point. We refer to the time increment between frames, measured in some unit of time, as the **frameshift**, and a larger frameshift implies a coarser/sparser **sampling resolution**, while a smaller frameshift implies a finer/denser sampling resolution. We pay special attention to sampling resolution in Chapter 4.

All these terms are discussed further in the thesis. We now situate and motivate the study of phonological maps for understanding how tones are acquired from the speech signal and more carefully define what we mean by phonological map.

1.1.2 Defining phonological maps

This thesis investigates the learnability of maps from the speech signal to lexical tonal concepts in tone languages. It is a preliminary step in the study of a much larger research question:

(Q0) *How do children acquire phonological categories from the speech signal?*

We address (Q0) using computational modeling, like previous studies of phonological category learning, e.g. de Boer and Kuhl (2003); Lin (2005); Toscano and McMurray (2010); Vallabha et al. (2007), and moreover, we ground our modeling assumptions based on phonetic fieldwork and perception experiments we conducted.

While a complete answer to Q0 necessitates a battery of behavioral, physiological, production, and perceptual studies on infants from the womb to adulthood, particularly in the first years of life, our ability to probe infant knowledge of phonological categories and connect this knowledge to their language input is limited, cf. methodological approaches in Polka et al. (1995); Werker et al. (1998). Thus, we make the choice to generalize our study to *any learner* so that we can deploy mathematically-specified learners to learn from examples we have very fine control over. The advantage of relying on computational approaches is that we can make a tight connection between the data that a learner may be exposed to (the domain of the learner, $\mathcal{D}$, the collection of all possible data sets), the hypothesis space of the learner (the range of the learner, $\mathcal{R}$, the set of all possible phonological categorizations), and the map from the collection of data sets to the hypothesis space of the learner (the learning function, $\mathcal{A}$). The challenge then is to also maintain a tight connection between the computational modeling and what we know about human learners.

Thus, we modify our original research question:

(Q0′) *How could a learner $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ from the collection of all possible data sets, $\mathcal{D}$, to the set of all possible phonological categorizations, $\mathcal{R}$, acquire lexical tonal categories from the speech signal in a way consistent with our knowledge about how humans do it?*

A key component in maintaining a tight connection between the computational modeling and human cognition is to have a clear picture of what the target of learning is (Dyson, 2004; Minsky and Papert, 1971). Thus, the goal of this thesis is to bear on the definition of the range of the learner, $\mathcal{R}$, the hypothesis space of the learner in the acquisition of lexical tonal categories: we study the hypothesis space to make progress towards characterizing if the target of learning in the acquisition of lexical tonal categories in natural language is feasibly learnable, as outlined in the overview.

What does it mean to have learned the tones of a tone language, e.g. the four basic tones of Mandarin: Tones 1-4, respectively, ˥ (high level), ˧˥ (rise), ˨˩˦ (fall-rise), ˥˩ (fall)? It means that a learner has learned a map from some subset of the data sets, D, out of the collection of possible data sets $\mathcal{D}$ (for Mandarin) to a subset of *tonal maps*, R, (for Mandarin) in the set of possible tonal maps in the range (hypothesis space) of the learner, $\mathcal{R}$,[1] as is made (redundantly) explicit in re-expressing $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ as the equivalent:

[1] A tonal language speaker does not learn one particular static mapping, between one particular data set and one particular tonal map. Learners are learners for life, so that their data set is continuously updated and their tonal maps as well.

$$\mathcal{A} : \{D | D \in \mathcal{D}\} \rightarrow \{R | R \in \mathcal{R}\} \tag{1.1}$$

where, under the methodological abstraction taken in this thesis to restrict the data to phonetic data (§1.1.4), tonal maps are a subset of a more general kind of map, which we call a *phonological map*, with the working definition:

$$\textit{Phonological map:} \quad \{\text{physical speech signal events}\} \rightarrow \{\text{phonological categories}\} \tag{1.2}$$

where the phonological categories are lexical tonal categories for tonal maps. Restricting the data $\mathcal{D}$ to the learner to come from the physical speech signal event—excluding, for instance, the context of the real-world situation in which the event occurs—means that in our model of the learning problem, examples for the learner are drawn from some phonetic space. This phonetic space is the domain of the phonological map.

We show a familiar example of a well-studied phonological map in Fig. 1.1, a vowel formant plot (Peterson and Barney, 1952). This is a map in a two-dimensional phonetic space $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ (over the steady-state values of the first and second formants) which maps unit-length sequences of phonetic parameter vectors $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ to English vowel phonemes, cf. Table 1.

There are some things to note from Fig. 1.1 and Table 1.1 which are general properties of phonological maps:

1. The domain of the phonological map is specified as sequences of phonetic parameter vectors rather than as physical speech signal events as in Definition 1.2. We emphasize that the parameterization for the domain consists

$\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$	English vowel phoneme	Note
$\langle 240, 2280 \rangle$	{/i/}	Actual data point
$\langle 460, 1330 \rangle$	{/ɝ/}	Actual data point
$\langle 475, 1220 \rangle$	{/ʊ/}	Actual data point
$\langle 686, 1028 \rangle$	{/ɑ, ɔ/}	Ambiguity
$\langle 400, 3500 \rangle$	{/i/}	Not a data point
:	:	

Table 1.1: The phonological map from steady state $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ formant space to English vowel phonemic categories from Peterson and Barney (1952). Formant measurements are from values reported in Praat (Boersma and Weenink, 2010) using the command `Create formant table (Peterson & Barney 1952)`.

of *sequences* of vectors because the speech signal unfolds in time and thus we must consider the extraction of parameters over time. Expressing the events with a lossy parameterization is in practice unavoidable—clearly encoding vowels as two steady-state formant values is lossy; even standard higher-dimensional parameterizations in automatic speech recognition such as a set of 39-dimensional mel frequency cepstrum coefficient vectors extracted with frameshifts on the order of 10 ms are lossy. What the parameterization should be is a choice the scientist must make, with an infinite number of choices available. What the appropriate phonetic parameterization of the physical speech event is, though, is hidden to us and needs to be determined under some decision criteria.[2] Making this determination is non-trivial, but it is necessary for us to posit some finite parameteri-

[2] The meta-cognitive inquiry readily available to humans for understanding the communication of morphosyntactic meaning seems to be unavailable to us for understanding speech communication at the physical and perceptual level of the speech signal.

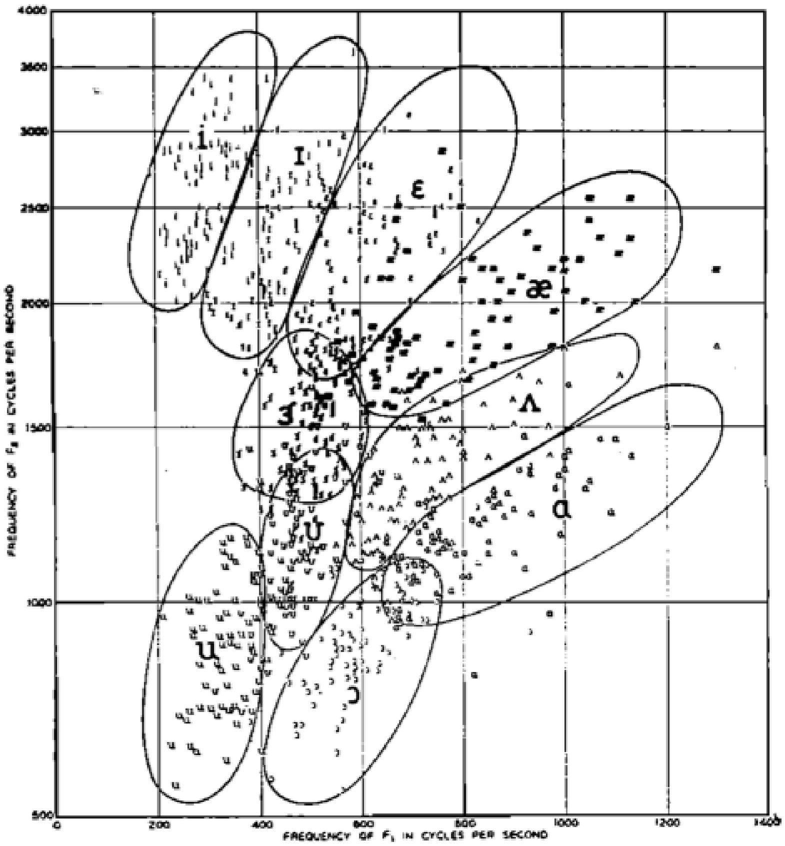


FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

Figure 1.1: A classic example of a well-studied phonological map, the vowel formant plot (Peterson and Barney, 1952).

zation to get a handle on the learning problem: the appropriate phonetic parameterization for tonal maps is, in fact, the main focus of this thesis. Without such a parameterization, it is impossible to represent the data to the learner in a computational model that is relevant for understanding the human learning problem. Without such a parameterization, it is also impossible to have the requisite scientific understanding of what the structure of the class of possible learned phonological maps is to be able to characterize the learnability of phonological maps, since the structure is conditioned on the parameterization.

One further note about the parameterization of the domain of phonological maps exemplified by the Figure 1.1 is that the phonetic parameter vectors are physical, acoustic parameters that can be measured from recordings. As we discuss in §1.1.4, the map of Definition 1.2 might be expanded as a chain of intermediate maps, for instance, including one from the physical parameters to ones associated with auditory perception. Because the physical parameters rather than the parameters internal to the brain and mind are the ones most accessible to us both by measurement and in our understanding of them, we generally restrict our exploration of phonetic parameter spaces to physical ones in this thesis.

2. There are regions of $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ space where the same $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ point is mapped to multiple English vowel phonemes: regions where vowel ellipses overlap. This highlights that *ambiguity in phonetic-phonological maps implies a range of sets of phonological categories rather than of single phonological categories.* From this observation and the previous one on parameterization, we revise our definition of phonological maps:

$$\begin{aligned} \textit{Phonological map:} \quad & \{\text{sequences of phonetic parameter vectors}\} \rightarrow \\ & \qquad \{\text{sets of phonological categories}\} \end{aligned} \tag{1.3}$$

3. The map is not total: there are some points in Figure 1.1 that are not enclosed in any of the regions labeled by a vowel category. This may also be the case for the true map being modeled and for phonological maps in general: there may exist physical speech events which are not mapped to any phonological category. While the map in Figure 1.1 is not total,

> the phonological vowel categories are each mapped to simply connected bounded and continuous regions over the phonetic space, and the regions are roughly ellipses in shape. Thus, each region includes infinitely many points that are not in the learner's finite data sample, and the map represents generalization in the learner from finite sets of examples of each category to categories consisting of non-denumerably infinite sets of points.

As Pierrehumbert (1990) discusses, phonological maps have parallels to "semantic" maps in morphosyntax:

$$\text{Morphosyntax} : \{\text{sequences of morphemes}\} \rightarrow \{\text{sets of meanings}\} \tag{1.4}$$

There is ambiguity in form-meaning mappings in morphosyntax, too, especially when we abstract away from relevant context (e.g. pragmatic and prosodic context in morphosyntax maps; morphosyntactic context in phonological maps); moreover, note that generalization from a finite data sample to an infinite language occurs for both learning problems. One difference between the phonological and morphosyntax maps is that phonological maps are defined in the real rather than the discrete domain.[3] Because of this, the mathematical machinery for studying the two different kinds of maps can differ.[4]

Having explicated an example of a phonological map, we make a final important revision to our current working definition in (1.3), after Pierrehumbert (2003a): we revise the range of the map to be over probabilities of category membership rather than sets of categories:

[3] There are also approaches to studying morphosyntax that model morphosyntax maps as being real-valued, cf. Widdows (2004): the co-occurrence of words in documents is used to determine similarity of word meanings, measured in real-valued vector spaces.

[4] It is possible to define a discrete phonological map. In fact, one may argue that a phonological map is most correctly modeled over a discrete space because of precision limits in computing and biological systems (Blum, 2004; Blum et al., 1997).

$$\textit{Phonological map:} \quad \{\text{sequences of phonetic parameter vectors}\} \rightarrow P_1 \times P_2 \times \cdots \times P_c \tag{1.5}$$

where each $P_i \in [0, 1]$ is the probability that the vector is an instance of phonological category C_i and there are c categories in total.

As an example for a particular mapping for vowels in 2-D formant space, we move from classification into a set of phonological categories which pick out subsets of $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ space:

$$\langle \mathit{F1}_{SS} = 686, \mathit{F2}_{SS} = 1028 \rangle \mapsto \{\text{/ɑ, ɔ/}\} \tag{1.6}$$

to a probability distribution of the categories over $\langle \mathit{F1}_{SS}, \mathit{F2}_{SS} \rangle$ space:

$$\langle \mathit{F1}_{SS} = 686, \mathit{F2}_{SS} = 1028 \rangle \mapsto \{p(\text{/ɑ/}) = 0.45,\ p(\text{/ɔ/}) = 0.55\} \tag{1.7}$$

Having defined the object of study for the thesis—phonological maps, and in particular, tonal maps—we describe our strategy for studying these maps to progress towards an assessment of the learnability of tonal maps in human languages.

1.1.3 Research questions

The initial step this thesis takes towards studying the acquisition of tonal maps is to focus on studying the hypothesis space of the learner because we are interested in characterizing the structure of the hypothesis space to assess if the class of tonal maps is feasibly learnable. Our strategy is to study properties of tonal maps in a diverse sample of languages. These maps for particular languages give us an indication of what possible tonal maps for human languages are. Based on what we learn about the structure of the class of possible tonal maps from the maps studied in the thesis, we then consider the learnability of this class.

The research questions we need to address to characterize tonal maps immediately follow from the definition of phonological maps in (1.5). We may frame the explication of these questions in terms of tonal maps due to the topic of the thesis, but the questions are relevant for the study for any kind of phonological map.

(Q1a) *What kinds of phonological categories are to be represented in the range of the map?*

(Q1b) *What is the phonetic parameter space—the space of phonetic parameter vectors—for the phonological categories defined in (Q1a)?*

(Q1c) *What are properties of the distributions of the phonological categories of (Q1a) over the phonetic parameter space of (Q1b)?*

Phonological categories (Q1a) The choice of definition for the categories referenced in the range of the phonological map revolves around how contexualized the categories are. Peperkamp et al. (2006); Pierrehumbert (2003a,b) argue for the set to be a set of positional allophones, and for unification into phonemes using information from distributions of symbolic allophones or by using knowledge of the lexicon; Dillon et al. (Unpublished) argues for the set to be phonemes.

Another option is to define the codomain over phonological features (Lin and Mielke, 2008; Mielke, 2008). We begin by restricting attention to tonal phonemes as the phonological categories in this thesis.

Phonetic parameter spaces (Q1b) We characterize the domain of the phonological map by motivating which phonetic parameters are most significant for defining phonological categories; these are the dimensions that we want to define the distributions over to make progress towards understanding the structure of the class of possible tonal maps. The set of phonetic parameters that may be extracted from the speech signal is obviously infinite in size and therefore must be constrained by some metric for computational tractability. For scientific purposes, too, we seek to limit the dimensionality of the phonological map, i.e., the size of the parameter set, in order to have a succinct representation that is intelligible to the human scientist (Occam's razor). From the learner's perspective, a succinct learning target prevents overfitting to the input data and facilitates generalization to novel data (Duda et al., 2001, 8-10), (MacKay, 2003, 343–349); from our scientific perspective, a succinct characterization of phonological maps facilitates our ability to understand how the learning proceeds. In the best case, succinctness in the map results in no loss of information, i.e. without any smoothing out of the distributional modes corresponding to category structure in the phonetic space;[5] otherwise, the goal is succinctness with minimal loss of information.

Our decision criterion for the problem of determining the appropriate parameterization of speech signal events in tonal maps, first introduced in the explication

[5] A simple example of succinctness without information loss is the expression of a finite language as a finite state automaton rather than as a list, since it can take fewer symbols to specify the finite state automaton than the list, and exactly and only the same sentences in the language are expressed (Meyer and Fischer, 1971).

of phonological maps in the previous section, §1.1.2, is therefore based on *the separability of tonal categories in a particular phonetic parameter space.*

We are in fact interested in characterizing three classes of phonetic parameter spaces to answer (Q0$'$):

1. a universal parameter space $\mathcal{U}$ for all tone languages
2. the language-specific parameter space $\mathcal{L}$ for a given tone language
3. the speaker-specific parameter space $\mathcal{S}_{\mathcal{L}}$ for a given speaker of a given tone language.

To review, by a parameter space, we mean the set of parameters over which the space is defined; the inclusion of a parameter adds a dimension to the space. By universal parameter space, we mean the smallest universal parameter space, the space which includes exactly and only the union of all language-specific parameter spaces.[6]

To a first approximation, we assume:

$$\forall \mathcal{L}, \forall \mathcal{S}_{\mathcal{L}},\ \mathcal{U} \supseteq \mathcal{L} \supseteq \mathcal{S}_{\mathcal{L}}. \tag{1.8}$$

This entails that the universal parameter space $\mathcal{U}$ can draw more distinctions than any tonal language-specific parameter space $\mathcal{L}$, which can, in turn, draw more distinctions than any speaker-specific parameter space for that language, $\mathcal{S}_{\mathcal{L}}$.

[6] The notion of a parameter space for all tone languages assumes that the class of tone languages is definable as a subset of all natural languages. Whether the restriction of languages to tone languages is available in acquisition is an open question, i.e. do children know they are learning a tone language, and if they do, under what conditions do they do this, and how do they do this?

The assumption in (1.8) is motivated by the overarching idea based on empirical work on infant speech perception development over the past few decades that infants begin as "citizens of the world" in having a universal ability to distinguish between sound categories and develop language-specific maps of the acoustic space through exposure to language input (Kuhl, 2004). For instance, one of the first results of this kind was that English-learning infants showed behavioral responses consistent with the ability to discriminate between a velar stop (a sound in English) and a uvular stop (a sound not in English, but in Salish) at 6-8 months of age, but that by 10-12 months of age, they did not anymore (Werker and Tees, 1984). Subsequent work confirmed and built on these results to flesh out a developmental timeline of perceptual reorganization of the acoustic space in which:

- Infants show a decline in their ability to discriminate nonnative vowel contrasts between 4-6 months (e.g. Polka and Werker, 1994).
- Infants learning a non-tonal language show a decline in their ability to discriminate lexical tonal contrasts between 6 and 9 months (Mattock et al., 2008).
- Infants show a decline in their ability to discriminate nonnative consonantal contrasts between 6-8 and 10-12 months (e.g Werker and Tees, 1984).
- Infants show improvement (facilitation) in their ability to discriminate native consonantal contrasts over the first years of life (Kuhl et al., 2006; Sundara et al., 2006).
- Infants may be able to discriminate some native contrasts only after exposure to native language input[7] (Narayan et al., 2010).

[7] Based on results like these, an alternative assumption to (1.8) is that the universal param-

- A nonnative contrast that infants show a decline in discriminating can be learned by adult speakers of the same native language after significant exposure to the nonnative language (Tees and Werker, 1984).

The cross-linguistic variability in the dimensions of acoustic spaces for phonological contrast and distributions of phonological categories over these spaces, as well as the change in the dimensions and distributions for infants due to language input show that *phonological maps must be learned from language input*: this has far-reaching ramifications for the status of the universality of phonological features (defined over phonetic parameters) and thus also of the universality of phonological constraints (referring to phonological features). It also underscores the need to study phonological maps using cross-linguistic data to answer (Q0$'$) on page 10 and is a kind of reiteration of the fact that the learnability of phonological maps can only be studied when we consider the structure of class of possible phonological maps in human language rather than phonological maps in a single particular language.

The empirical evidence that: (i) language learners show decline rather than loss in sensitivity to particular phonetic dimensions, (ii) they can reactivate sensitivity with later language exposure and training, and (iii) listeners show the ability to use a wide variety of cues in degraded speech[8] suggests that the model

eter space $\mathcal{U}$ can draw *fewer* rather than *more* distinctions than any tonal language-specific parameter space $\mathcal{L}$:

$$\forall \mathcal{L},\ \mathcal{L} \supseteq \mathcal{U}. \tag{1.9}$$

following the idea that sensitivity to some phonetic parameters may become activated only after exposure to language input. We do not take this alternative assumption because there is, to date, little supportive evidence for it. More importantly, a negative result for infant sensitivity to a speech sound contrast is conditional on a given experiment using a given task. A positive result is conditioned in the same way, as well, but shows that, at least under some conditions, infants show sensitivity to the contrast, while a negative result does not imply that infants are not sensitive to the contrast under any conditions.

[8] See Assmann and Summerfield (2004) for a general review of perception of degraded speech.

of the development of language- and speaker-specific spaces of each language involves *parameter tuning/re-weighting* rather than *parameter selection*. Even in cases where sensitivity to some phonetic parameter may be vanishingly small, the model should assign it a vanishingly small weight rather than remove the parameter from the space.

Note that even for the purposes of studying the phonetic parameter space, we must represent data with a set of initial parameters: this initial set should be exactly $\mathcal{U}$, which we assume to be a superset of the dimensions of $\mathcal{L}$ for any natural tone language $\mathcal{L}$, cf. (1.8), and which is a subset of the set of all acoustic parameters we could extract from the speech signal. But these are not well-defined lower and upper bounds on $\mathcal{U}$; we cannot know what $\mathcal{U}$ is before studying what it should be! Thus, we make a guess and initialize the parameter set of $\mathcal{U}$ based on cross-linguistic work on tonal production, perception, and automatic tonal recognition.

The distribution (Q1c) We assume that the distribution of phonological categories over the phonetic space is continuous. Since the details of the distribution depends strongly on how the phonological categories and the phonetic space is defined, we let our study of those determine characteristics of the distribution. These characteristics then inform how we constrain the type of distributions available in the hypothesis space for the learner in modeling the actual learning of the phonological map—the structure of the distributions in the class of possible phonological maps is what informs us about the learnability of the class.

(Q1a)–(Q1c), reiterated below, are questions that we ask of *particular* tonal maps.

(Q1a) *What kinds of phonological categories are to be represented in the range of the map?*

(Q1b) *What is the phonetic parameter space—the space of phonetic parameter vectors—for the phonological categories defined in (Q1a)?*

(Q1c) *What are properties of the distributions of the phonological categories of (Q1a) over the phonetic parameter space of (Q1b)?*

The learnability of the hypothesis space (Q1d) To these, we add (Q1d), a question about the class of possible tonal maps which we address after studying tonal maps in a range of particular languages:

(Q1d) *Considering the tonal maps of the languages studied as an indication of the possible maps in human language, is the class of possible tonal maps feasibly learnable?*

As we mentioned in the overview, our definition of learnability is based on a mathematically precise criterion for successful learning. **The criterion we take for learnability—more specifically, *feasible* learnability—is that the hypothesis space of the learner must have finite Vapnik-Chervonenkis (VC) dimension**, a combinatorial measure of the complexity of the hypothesis space (Vapnik and Chervonenkis, 1971). VC dimension is often described in an intuitive way as reflecting the rigidity or flexibility/wiggliness in the structure of the hypothesis space (and in the capacity of learners for the hypothesis space). For instance, the VC dimension of the class of ellipses in $\mathbb{R}^2$, like the approximate shape of vowel concepts in Figure 1.1, is small and finite, while the VC dimension of the class of convex polygons (shapes including triangles, quadrilaterals, ellipses...) in $\mathbb{R}^2$ is infinite (Blumer et al., 1989). VC dimension, as a combinatorial measure, can be calculated in discrete as well as real-valued spaces, and thus serves as a unified metric for learnability in language for classes in phonology and morphosyntax, e.g. context-free and context-sensitive languages, as well as for concept classes in phonetics spaces such as tonal maps.

The criterion of finite VC dimension is a necessary and sufficient criterion for probably approximately correct (PAC) learnability (Blumer et al., 1989) (and more generally a class of learnability criteria including PAC learnability (Poggio et al., 2004)). PAC learnability is a probabilistic criterion on successful learning that does not demand exact convergence on the target, and thus is considered to be a learnability criterion more relevant for human learning than stricter criteria. It requires only that for a sufficiently large sample of data drawn according to some distribution, the learner converge within arbitrarily small error with arbitrarily high probability on any target map for all maps in the hypothesis space (Valiant, 1984).

Finite VC dimension is a property of the hypothesis space. Studying learnability from the perspective of classes of algorithms may be a fruitful strategy if we have a good understanding of the learning algorithms for the learning problem in question (Poggio et al., 2004), but we focus on learnability from the perspective of structure in the hypothesis space because we know more about the tonal maps to be learned than the learning algorithms for them at this point in time.

1.1.4 Methodological abstractions

The focus of this thesis is working toward answering (Q1b), and in characterizing the phonetic parameter space for tonal maps, we make four main methodological abstractions, some of which we have already mentioned: (i) to sharpen the probabilistic distributions of phonological categories into partitions over the phonetic space, (ii) to use category separability as a step towards constraining the phonetic parameter space, (iii) to limit the context available for phonetic parameter extraction from the speech signal, and (iv) to introduce linguistic structure into the unanalyzed speech signal. All of these abstractions are in addition to the overarching abstraction to restrict the set of languages under study to tonal languages, as described in Footnote 6 on pg. 19. Characterizing the phonetic parameter space with these methodological abstractions in place still allows us to bear on questions (Q1a)–(Q1d). We discuss each abstraction below.

Partitions over the phonetic space and category separability While the reality is that phonological maps are probabilistic distributions of phonological categories over the phonetic space, in characterizing the phonetic parameter space, we make the methodological abstraction that maps are *partitions* of phonological categories over the space: every point in the space maps to exactly and only one phonological category.

The reason for the abstraction is that most well-understood computational algorithms for classification give "hard" classifications, i.e. produce a partition of the space, rather than a probabilistic distribution over it (Wahba, 2002). Moreover, while it is possible to elicit probabilistic confidence ratings in human perception experiments, e.g. using magnitude estimation (Bard et al., 1996; Keller, 2000), we use forced choice tasks in our perception experiments to match the hard classification of the computational algorithms.

Along with the methodological abstraction of modeling phonological maps with partitions, we use the an assessment of *category separability* to determine how relevant/informative phonetic parameters are for defining the tonal categories in computational modeling: more informative phonetic parameters define a space in which the tonal categories are better separated. As discussed by Nearey (1989), this category separability metric is *data analytic* because it is based on production data only, while ultimately, *perceptual* separability from listening experiments, as well as *articulatory* separability from physiological experiments, are also directly relevant for defining phonological maps. However, data analytic category separability certainly bears on perceptual separability, and at this point in time, we can make sharper conclusions from a data analytic perspective than a perceptual one because we have a better understanding of the spaces used in a data analytic approach.

Limiting context for phonetic parametrization We have already proposed (1.2) restricting the domain of the maps to be learned to phonetic parameters. We reiterate here that we are abstracting away from non-phonetic context, e.g. morphosyntactic information (the language model in automatic speech recognition), to constrain the research problem; Jansen (2008) calls this the "pure speech" setting. Moreover, we restrict the *temporal domain* for phonetic parameter extraction. The strongest such restriction is to restrict the extraction of phonetic parameters to only the unit to the classified, e.g. only from the syllable of the tone to be classified. Unless otherwise specified in the thesis, we start from this restriction, and in the immediately following chapter, Ch. 2, we study the effect of that restriction and allow extraction from the preceding and following syllables as well. For fluent speech recognition by humans, there is strong evidence that humans extract parameters from temporal domains wider than the unit to be classified, e.g. Ladefoged and Broadbent (1957); Wong and Diehl (2003).

Introducing linguistic structure in the speech signal While the original research question (Q0′) assumes extraction of parameters from the unanalyzed signal, for this paper, we extract parameters from speech segmented for syllabic structure for convenience. This is like having an oracle tell the classifier where syllable boundaries or onset/rime boundaries are. In future work, we can remove this extra information by implementing a sonority detector to find syllables, as in Jansen (2008).

1.2 The model system for the acquisition of lexical tones

With the larger research questions and the methodological abstractions set up, we turn to the model system under study.

The gross characterization of our model system for the thesis is this:

- **Data**: monotones, bitones, and tritones extracted from sentence-medial position in connected speech and in isolation over a cross-linguistic tonal language sample
- **Phonetic parameter space**: an acoustic parameter space, with parameters extracted from the speech signal, used to represent the data
- **Phonological categories**: lexical tonal phonemes (tonemes)

Like any other system studied in phonological category acquisition, the one we study here is a model system, and we study it with the same scientific motivation that a biologist studies a simple model organism like baker's yeast (the eukaryote with the smallest number of genes) to illuminate gene regulation in more complex systems such as humans (Fields and Johnston, 2005). Clearly the model system can only capture certain aspects of the process of phonological category acquisition, highlighting some while muting others (§1.3). In this section, we describe how we instantiate the model system for lexical tone acquisition to answer (Q0′).

Our research questions, as laid out in §1.1, dictate the following requirements for setting up a model system for studying lexical tone acquisition:

- A representative cross-linguistic sample to address the language-specific development of speech categorization and to reveal potential structure in the class of possible tonal maps for assessing learnability
- A language sample relevant for modeling language input to infants
- Some controlled source(s) of variability to enable modeling the challenge of categorization in the face of variability

Cross-linguistic tone language sample We chose a sample of tonal languages to include: (i) a register/level tone language, with only level tones (Bole), and (ii) contour tone languages with contour tones and level tones (Mandarin, Cantonese, Hmong), as well as languages with a variety of tone-voice quality interactions.

While Bole is not known to have any such interactions, both Mandarin and Cantonese have noncontrastive phonation, i.e. tone-voice quality interactions occur, but f0-based cues are thought to be sufficient to distinguish between all tones. Mandarin has creaky phonation particularly in its dipping tone (Hockett, 1947; Chao, 1956; Gårding et al., 1986; Klatt and Klatt, 1990; Davison, 1991; Belotel-Grenié and Grenié, 1997), and Cantonese has anecdotally been claimed to have a creaky low fall (Vance, 1977). Hmong has a laryngealized low tone and a breathy fall (Esposito et al., 2009; Andruski and Ratliff, 2000; Huffman, 1985). Moreover, in White Hmong, the dialect of Hmong we investigated, there is production evidence that breathy phonation is contrastive for female speakers: both the modal and breathy high fall can have very similar f0 contours and heights (Esposito et al., 2009).

We summarize the diversity of the cross-linguistic tonal language sample below in Table 1.2, using International Phonetic Alphabet notation for the tonal inventory, and give recording details of the data currently available below in Table 1.3.

In our perceptual experiments in Chs. 4, 2 and 3, we used Cantonese tones as our system of study. This was for two reasons. First, Cantonese contains level and contour contrasts, and thus can inform us about tonal representations for more tone languages than a language with only dynamic contrasts like Mandarin (level, rise, fall, dip) or Bole (level contrasts only). Second, because behavioral experiments require a ready supply of subjects, we chose Cantonese for practical reasons, as it is reasonably easy to find native speakers of Cantonese to participate in experiments.

Language input to infants and sources of variability Because infants tune into native language sound categories before they begin to produce native language sounds (Werker, 1994; Best, 2003; Kuhl, 2004), we restricted parameters to *acoustic parameters* and abstracted away from articulatory parameters as a

Language	Area	Tonal inventory	Phonation
Bole	Nigeria	˥, ˩ (H,L)	
Mandarin	Beijing, Taiwan	˥, ˧˥, ˨˩˦, ˥˩	creaky ˨˩˦, ˥˩
Cantonese	Hong Kong	˥, ˧, ˨, ˨˩, ˨˥, ˨˧	creaky ˨˩
Hmong	Laos/Thailand	˥, ˧, ˨, ˥˨, ˦˨, ˨˩, ˨˦	breathy ˦˨, creaky ˨˩

Table 1.2: Cross-linguistic sample of tonal languages recorded to provide language input

Language	Dialect	Recording location	Speakers
Bole	Fika	Potiskum, Nigeria	3M/2F
Mandarin	Beijing	Beijing, China	6M/6F
Cantonese	Hong Kong/Macau	Los Angeles, CA	6M/6F
Hmong	White	Fresno, CA	6M/5F

Table 1.3: Details for recordings of language sample

methodological abstraction.

Other work on learning tonal categories has emphasized that the majority of the input to the infant consists of multiple words so that contexual variation due to tonal coarticulation from neighboring tones is a regular part of the input the learner receives (Gauthier et al., 2007; Shi, in press). Specifically, Gauthier et al. (2007); Shi (in press) claim that about 90% of parental speech to infants is multi-word utterances. Moreover, the majority of language data an infant hears is not speech directed to the infant, but, for instance, adult-to-adult speech. An estimate from van de Weijer (1998, 2002) is that only about 14% of the input is direct speech to the infant.

Because of the large amount of input that infants hear that is adult directed speech and multi-word utterances, Gauthier et al. (2007) modeled learning tone categories based on speech from adults rather than infant-directed speech, (and in general, research building tone recognizers is modeled on adult speech). This is of course a working hypothesis; surely the presence of infant directed speech and isolated words in the input could affect the character of the learning problem.[9] We follow this choice, taking our input to the learner to be adult connected speech. We capture the role of contextual tonal variation in creating variability in the input by collecting the full permutation set of bitones in connected speech for each language in the sample, and we capture interspeaker variation by recording multiple speakers of both genders. We use language samples of this kind in all our perceptual experiments (Chs. 2, 3, and 4) and in our computational modeling (Chs. 2, 4, and 5).

1.2.1 Stimulus sets

The stimulus sets for each language were recorded as all possible bitone combinations in the language, in connected speech. For all languages, we attempted to elicit these bitone targets as CVCV (or CDCD, D for diphthong) sequences. For the African languages, the bitone targets were disyllabic words; for the Asian languages, they were (mostly) nonce disyllable sequences of identical CV/CD syllables. In order to produce smooth f0 tracks, we selected sonorant consonants whenever possible; in order to produce segmentable speech, we selected nasals

[9] For instance, note that the rationale for the ecological validity of adult connected speech given above assumes equal weighting in infant attention to all input regardless of whether it is directed to the infant. In fact, studies show biases for infant directed speech over adult speech and biases for the infant for their mother's voice and the importance of placing language input within social interaction (Kuhl et al., 2003). Thus, it is not unreasonable to hypothesize that despite the relatively small amount of infant directed speech in the ambient input, it may be a rich source of information for infants about learning tone patterns. In fact, work has found correlation between the amount of exaggeration in infant directed speech in terms of the expansion of the vowel and tonal spaces in predicting an infant's ability to discriminate native consonant contrasts (Liu et al., 2003; Xu and Burnham, submitted).

and laterals rather than liquids and glides, when possible (Xu, 1997). In order to facilitate reliable voice quality measures, we selected low vowels when possible, mostly [a], since when F1 becomes low enough to interact with f0, the voice quality measurements reliant on accurate formant estimation are not reliable. Within a language, with the exception of Igbo, only one vowel or diphthong was used in the stimuli to control for and abstract away from intrinsic f0 effects.

We recorded the bitones in sentence-medial position, flanked by surrounding tones. For languages with a large tonal inventory, such as Cantonese and Hmong, we kept these constant at the mid level tone in the inventory. In the other languages, we elicited the bitones in a variety of tonal frames, as detailed below.

The stimuli described below were used in cross-linguistic computational modeling, described in Ch. 5, and the Cantonese stimuli were used for perceptual experiments on creak in Cantonese tonal representation (Ch. 3), and very similar stimuli were also used for the other Cantonese perceptual experiments in Ch. 4 and 2.

1.2.1.1 Bole

Factor	Levels
Tone$_{S1}$	L, H
Tone$_{S2}$	L, H
Context	H__H, H__L, L__H, L__L
Prosodic pos.	Controlled as sentence-medial
Segmentals	Mixed vowels, mixed sonorant consonants
Other elicitation	Citation, isolated disyllables, monosyllables in context

Table 1.4: Factors in Bole corpus

Segments	Tones	Gloss
lala	HH	spider
mono	HL	cobra
yaro	LH	bird
nema	LL	prosperity

Table 1.5: Bole targets

- Contexts:
  - Left contexts: *nzono* LH ‘yesterday’, *tuwwa* HH ‘day after tomorrow’, *anin* HH ‘owners’
  - Right contexts: *mengo* HL ‘come back’, *mai* L ‘return’
- Morphosyntax: Subject DP after time adverb or object in genitive construction

1.2.1.2 Mandarin

Factor	Levels
Tone$_{S1}$	˥, ˧˥, ˨˩˦,˥˩
Tone$_{S2}$	˥, ˧˥, ˨˩˦, ˥˩, (and neutral tone)
Context	L__L, L__H, H__L, H__H
Prosodic pos.	Sentence-initial, medial, final
Segmentals	All targets [ma]

Table 1.6: Factors in Mandarin corpus

- Targets: disyllable combinations of all four tones on [ma]

- Contexts:
  - Sentence-initial: #__[jau]{˧˥,˥˩}
  - Sentence-medial: [jau]{˥,˥˩}__[mai]{˧˥,˥˩}
  - Sentence-final: [jau]{˥,˥˩}__#

1.2.1.3 Cantonese

Factor	Levels
Tone$_{S1}$	˥, ˧, ˨, ˨˩, ˨˥, ˨˧
Tone$_{S2}$	˥, ˧, ˨, ˨˩, ˨˥, ˨˧
Context	Controlled as ˧__˧
Prosodic pos.	Controlled as sentence-medial
Segmentals	All targets [lau]
Other elicitation	Citation, isolated disyllables

Table 1.7: Factors in Cantonese corpus

- Targets: disyllable combinations of all six tones on [lau]
- Contexts: [jɛu] ˧ __[jaak] ˧ (Mid checked)
  - Three instances of same tonal context frame: [jaak jœŋ] ‘eat sauce’, [jaak sou] ‘eat vegetarian’, [jaak gap] ‘eat pigeon’

1.2.1.4 White Hmong

- Targets: disyllable combinations of all 7 tones on [la]
- Contexts: *cia* __ *ya*

Factor	Levels
Tone$_{S1}$	˥, ˧, ˨, ˥˨, ˦˨, ˨˩, ˨˦
Tone$_{S2}$	˥, ˧, ˨, ˥˨, ˦˨, ˨˩, ˨˦
Context	Controlled as ˧__˧
Prosodic pos.	Controlled as sentence-medial
Segmentals	All targets [la]
Other elicitation	Citation, isolated disyllables

Table 1.8: Factors in White Hmong corpus

- Two instances of same tonal context frame (mid tones on left and right): *ya mus tov* 'fly there' and *ya mus tsev* 'fly home'

1.3 Reflections on the model system

We conclude this introductory chapter with reflections on the model system we have described. First, we consider limitations of the model system for understanding the acquisition of phonological categories §1.3.1. Then, we consider what properties of the phonetic parameterization of phonological maps we can investigate with our model system in comparison to previous model systems for learning phonological categories and describe which issues we focus on in this thesis in §1.3.2.

1.3.1 Limitations of the model system

There are many sources of variability in the acoustic realization of tone. The emphasis on contextual tonal variation in the collected data mostly restricts the source of variability in the phonetic realization of tones in our model system to tonal coarticulation. Because we recorded multiple speakers and genders, our model system also includes interspeaker variability. However, the interaction of tone and intonation is not emphasized in the data, other than implicitly in language-specific implementations of downtrend; we did not consider the effect of prosodically realized focus or other pragmatically-conditioned effects on f0; we did not systematically vary pitch range of an individual speaker, as in Liberman et al. (1992); Liberman and Pierrehumbert (1984); with our "pure speech" setting, we abstracted away from tonal interactions with morphophonology/syntax, including the rich grammatical function of tone in many tone languages. Moreover, the size of our speech corpora is tiny—at least one or two orders of magnitude smaller than speech corpora used in automatic speech recognition—and it consists of controlled laboratory speech rather than semi-spontaneous or spontaneous speech.

Because the variability in the input data recorded in our corpora may be a small subset of the variability characteristically present in input that an infant might hear, while the sample size of data given by our corpora is small compared to the amount of data available for inducing phonological categories in human learners, it is not possible to say whether analyses based on our corpora are likely to give an upper or lower bound on how well tonal categories are separated in a given phonetic space. How closely the settings of the learning problem in our work match that of the real situation is an open empirical question, but converging results on learnability from our work with multiple sources with different settings, e.g. from automatic tonal recognition, human tonal perception studies, and infant behavioral research would suggest that our model system provides a reasonable characterization of the learning problem.

1.3.2 The model system as a model system for phonological category acquisition

Our model system and methodology for modeling phonological category acquisition is unusual in several respects. First, we precede study of learning phonological maps with study of the maps themselves using *supervised learning* methods, where the phonological classification algorithm receives phonetic data labeled with which phonological category it is an instance of. Other phonological category learning studies, cf. Table 1.9, have bypassed this initial step and only studied the learning of the map using *unsupervised learning* methods, where the learning algorithm receives unlabeled phonetic data.[10]

Results from unsupervised learning simulations are obviously conditioned on how the learning problem is set up. As an illustration, in Bayesian models of phonological category acquisition, Bayes' rule for statistical inference for posterior beliefs about model parameter values after observing data (evidence), given informally as:

$$\mathit{posterior} = \frac{\mathit{likelihood} \times \mathit{prior}}{\mathit{evidence}} \tag{1.10}$$

and more formally, with data $\mathcal{D}$ and hypothesis θ for model parameter values as:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \tag{1.11}$$

can always be rewritten using the chain rule to make the conditioning of each term in the equation on the model of the learning problem $\mathcal{M}$ explicit:

[10] For vowel learning, there is a body of supervised learning work studying categorical separability conditioned on different phonetic parameter spaces, e.g. Hillenbrand et al. (1995); Hillenbrand and Gayvert (1993); Nearey (1992), but this has not been referenced by the unsupervised learning studies.

$$P(\theta|\mathcal{D}, \mathcal{M}) = \frac{P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} \tag{1.12}$$

Second, most studies of phonological category learning from the speech signal have focused on learning vowels (de Boer and Kuhl, 2003; Feldman et al., 2009; Dillon et al., Unpublished; Vallabha et al., 2007); some have modeled learning stop voicing (Toscano and McMurray, 2010) or an entire segmental phonemic inventory (Lin, 2005); only one has studied learning suprasegmentals, learning Mandarin tones (Gauthier et al., 2007) (Table 1.9).

Tones as phonological categories also have a different status from phonological categories such as vowels because tones are lexically contrastive, while vowels are sublexical, and the characterization of phonological feature systems for tones is not well understood, compared to that that for vowels and consonants (Hyman, 2010). Despite being an unusual system to study in phonological learning, our lexical tone model system allows the study of parameterization issues common to many phonetics-phonology maps (§1.3.2.1), as well as some particular to lexical tone maps (§1.3.2.2).

Study	Model system	Phonetic parameters
de Boer and Kuhl (2003)	English infant-directed /i, a, u/	F1/F2 tracks (1 sample/16 ms)
Dillon et al. (Unpublished)	Turkish front vowels, Inuktitut vowels	F1, F2, F3 from midpoint; steady state F1, F2
Vallabha et al. (2007)	Some infant-directed Eng./Jap. vowels	F1, F2 from first $\frac{1}{4}$/steady state, duration
Feldman et al. (2009)	English vowels	mean F1, F2 (Hillenbrand et al., 1995)
Toscano and McMurray (2010)	English voiced/voiceless stops	VOT, vowel length
Lin (2005)	English phonemes	MFCCs (1 sample/10 ms)
Gauthier et al. (2007)	Mandarin tones	f0/ f0 velocity tracks (28-30 samples/syllable)

Table 1.9: Model systems and phonetic parameters in studies of unsupervised phonological category learning

1.3.2.1 Common parameterization issues shared by lexical tone and other systems

In this section, we discuss parameterization issues common to many model systems of phonological maps that we can study with our model system.[11] Following Nearey (1989), we distinguish three classes of properties: (i) static, (ii), dynamic, and (iii) relational, and further subdivide relational properties into intrinsic (intrasegmental) and extrinsic (transsegmental) properties. For any model system, the relative contributions of these different classes of properties to category separability is of interest. For vowel systems, Nearey (1989) suggests that all three classes of properties contribute to separability and perception. Of special attention in this thesis are two types of properties: contextual temporal domain and sampling resolution, studied in Chs. 2 and 4, respectively.

Static properties *Static properties* are properties such as steady-state formant frequencies, extracted from the speech signal from a region where their rate of change is vanishingly small. They could also include mean values over the unit of speech presented as a data example, e.g., mean f0 over the syllable, mean F1 over the vowel, or spectral mean over the fricative (Jongman et al., 2000). Note that about half of the studies in Table 1.9 include only static properties in the parameterization of the speech signal.

Dynamic properties With the exception of Ch. 5, we confine our parameters to dynamic properties in the thesis. Dynamic properties include change in formant frequencies over a vowel (vowel inherent spectral change, VISC (Nearey and Assmann, 1986)), change in f0 over a tone, or change in spectral mean or

[11] Consonantal model systems have some nonoverlapping parameterization issues due to the transient character of consonantal acoustic realization.

kurtosis over a fricative, as well as coarticulatory effects from neighboring units of speech, such as contextual effects on formant frequencies from consonants and contextual tonal effects from neighboring tones. Dynamic properties are in fact a special class of relational properties (other classes are described next) where information is spread across time, across multiple time points/time slices; they are defined by sampling of timecourses, either within the unit of speech, or extending to neighboring units of speech for contextual effects. In this thesis, the dynamic properties we use in computational modeling are mean f0 values extracted over time slices uniformly over the syllable and the changes between these.

Intrinsic relational properties *Intrinsic relational properties* are properties derived from quantities measured at the same timepoint or averaged over the same time slice within the speech unit, e.g. within the vowel, or within the tone, such as steady state formant ratios or spectral balance measures, e.g. $H1 - H2$ (the difference between the amplitude of the fundamental frequency spectral peak and the amplitude of the second harmonic peak, taken at a timepoint or averaged over a time slice). We do not use these kinds of parameters in this thesis.

Extrinsic relational properties However, we do study *extrinsic relational properties*, which are quantities measured for a speech unit that are relativized using an overall frame of reference such as the ensemble of all formant measurements from a given speaker, such as relative formant values or relative f0 within a speaker's pitch range. Extrinsic relational properties include parameterization under normalization procedures: "explicit methods that attempt to factor out systematic, but phonetically nondistinctive, covariation in signal properties, and thus to reveal more invariant patterns separating phonetic categories" (Nearey, 1989, 2090). Such procedures include the Chao tone value notation system of five relative pitch levels within a given pitch range (Chao, 1930), or more generally, z-score normalization of frequencies based on the mean and variability over some ensemble of measurements (Rose, 1987), e.g. over the speaker or a particular vowel or tone category. Whether normalization procedures have a psychological basis or not is an open empirical question. In this thesis, preprocessing for computational modeling in Chs. 4 and 2 involves z-score normalization based on speaker-specific means and standard deviations. The misfit between human behavior directly observed for the same stimuli does not provide evidence for a psychological basis for z-score normalization of this kind in humans.

Similar to normalization transforms, *distance metric transforms* are transforms that do not rely on a particular frame of reference as for extrinsic relational properties; rather, they involve a transform in the scaling of distances. Common nonlinear distance metric transforms are the log and perceptually-based log-like distance metric transforms (erbs, semitones, bark); the effect of these transforms on category separability is complicated and interacts with the classifier algorithm specification (Hillenbrand and Gayvert, 1993). In this thesis, we do not study this, but assume that log-transformed f0 values are perceptually motivated and use them in computational modeling.

Contextual temporal domain The specification of dynamic properties extending to neighboring units of speech for contextual effects requires parameter extraction from those neighboring units. The extent of the *contextual temporal domain* sets an upper bound or "window size" for parameter extraction. From studies such as Ladefoged and Broadbent (1957); Wong and Diehl (2003) showing how preceding context biases perception for vowels and level tones, it is clear that the contextual temporal domain must extend to neighboring units, although the studies in Table 1.9 on page 39 abstract away from this extension. We study the contextual temporal domain in tonal representation in humans and in machine classifiers in detail in Ch. 2, using Cantonese tones as a case study.

Sampling resolution and related properties Multiple samples from timecourses are needed for the parametrization of dynamic properties, e.g. samples from the timecourse for the first formant for dynamic information about F1. Three properties controlling the sampling are the *sampling resolution*, clock, and adaptiveness. The sampling resolution controls how sparse or dense the sampling is while the clock and adaptiveness of sampling control the sampling rhythm. Sampling with a clock running on absolute time means sampling at some resolution at fixed intervals milliseconds apart; sampling with a clock running on syllable time means sampling at some resolution at fixed syllable fraction intervals, e.g. every tenth of a syllable. The sampling also need not be regular based on the clock, but could be adaptive: denser during regions of rapid change, or occuring only at landmarks, such as turning points (maxima, minima) in the timecourse (Stevens, 2002). These temporal properties are generally not discussed in the linguistic literature, which focuses on static properties, but they are sometimes discussed in automatic speech recognition (Jansen and Niyogi, to appear; Tian et al., 2004). Note that the studies in Table 1.9 including dynamic information in the representation all employ dense, fixed-rate sampling. We study temporal resolution in tonal representations in detail in Ch. 4. We assume uniform, non-adaptive sampling on a syllable-timed clock, a methodological abstraction, and consider the effect of sampling resolution on the separability of tonal categories in humans and machine classifiers, using Cantonese tones for a case study.

1.3.2.2 Representational issues special to lexical tone

Lexical tone systems also bring other representational issues to the forefront: one is a reflex of how tone is produced articulatorily and the other is a reflex of the temporal properties of tonal representation.

The voice source Unlike any of the other systems of contrast studied in Table 1.9, tone contrasts are regulated largely independently through the *voice source*, the glottis, according to the source-filter model of Fant (1960). In contrast, vowel and consonant contrasts, other than voicing contrasts and contrasts involving pitch such as in Korean stops, are, to a first approximation, regulated independently through the manipulation of vocal tract resonances *filtering* the voice source (Fant, 1960).[12] An example of this near-independence is that formant trajectories give resonance information (about the filter), while f0 tracks give periodicity information (about the source). As was implicit in §1.2, the consequence of tonal contrasts being regulated through the voice source is that *phonetic spaces for tonal maps involve other source parameters than fundamental frequency, which is just one property of voice quality.* However, the tonal study Gauthier et al. (2007) listed in Table 1.9 represents tones without voice quality information other than f0, and the same situation is true in automatic speech recognition (other than amplitude-related parameters), with the exception of Surendran and Levow (2008)'s work on tonal recognition using voice quality parameters.

Local temporal domain For segmental phonological categories, the unit of speech is easily defined as the temporal extent of the segment: a vowel is locally

[12] However, see Esling (2005) for an example of work on the laryngeal specification of vowels.

defined over its own temporal extent. However, the local temporal domain for the definition of tones is not as straightforward, and there are different assumptions that can be made regarding it. Generally, it is assumed that the domain is no larger than a syllable. Other assumptions for the domain account for the fact that there is no f0 information in voiceless regions of the speech signal, and that such voiceless regions may be present in the onset of a syllable; thus, other assumptions for the local domain are the rime, the vowel, or voiced regions, cf. Table 1.10. For this thesis, we used data from all-sonorant syllables and assumed a local temporal domain of the syllable.

Study	Language	Clock	Sampling resolution
Zhang et al. (2004)	Mandarin	Absolute time	Fine, 10ms constant frame rate
Gauthier et al. (2007)	Mandarin	Normalized to syllable	Fine, 30 samples/syll
Ọdẹ́lọbí (2008)	Yoruba	Normalized to syllable	Medium, 9 slices/syll
Wang and Levow (2008)	Mandarin	Normalized to tone nucleus	Coarse, 5 samples/ tone nucleus
Qian et al. (2007)	Cantonese	Normalized to rime	Coarse, 3 slices/final
Zhou et al. (2008)	Mandarin	Normalized to nucleus	Coarse, 3 slices/nucleus

Table 1.10: Sampling characteristics of a selection of tone recognition studies

For the studies in Table 1.9, the fact that the unsupervised learner implemented was able to learn the phonetics-phonology map of the model system is an existence proof that the learning target, e.g. the vowel system, is learnable in some sense. However, the settings for the model systems for these studies render them unable to account for well-known facts of human perception. For instance, all the model systems employing purely static steady state/mean formant parameters do not capture the fact that humans are able to identify vowels with only two short initial and final portions from the vowel as well as vowels with only the steady state central vowel region present (Strange et al., 1983). Similarly, if Poeppel (2003) is right that neuronal mechanisms chunk the speech signal into 20-40 ms and 150-250 ms temporal integration windows, the model systems based on densely sampled timecourses oversample relative to the maximum temporal resolution relevant for human cognition. By considering the parameterization issues discussed in this section by using human perception experiments in Cantonese as a case study and computational modeling in a range of tonal languages, we can situate models of learning tonal maps to be maximally relevant for understanding how humans learn them.

CHAPTER 2

Temporal domain in tonal representations: a case study with Cantonese tonal perception

2.1 Introduction

The classic phenomena we study to understand the nature of phonological representations in language are allophonic variations—the variability in the realization of a phonological form due to contextual influences—and alternation—the relations between these variants (Kenstowicz and Kisseberth, 1979; Hayes, 2008). This chapter addresses both these issues for the phonological representation of lexical tones from a perspective informed by automatic tonal recognition, using evidence from a human tonal perception experiment and computational modeling in Cantonese.

To recognize the tone associated with a syllable in the face of allophonic variation, automatic tonal recognizers that operate on connected speech are regularly given access to acoustic parameters from neighboring syllables (Chen and Wang, 1995; Zhang and Hirose, 2000; Zhang et al., 2000; Levow, 2005; Peng and Wang, 2005; Qian et al., 2007; Surendran, 2007). In addition, in automatic speech recognition for Chinese languages, the citation form of a tone—its form uttered in an isolated monosyllable (Chen, 2000, 49)—is often treated as definitional and implicitly as a privileged base from which other contextual variants of the tone are derived. The tonal inventories of Chinese languages are typically defined over citation forms, using iconic or 5-level Chao tone letters (Chao, 1930), and Mandarin Tone 4 is standardly defined as a fall, as it appears in isolation, even though there are contexts in which it may appear as a rise (Shih and Kochanski, 2000).

Two examples of Mandarin automatic tonal recognition systems that implicitly treat citation forms as privileged bases are the Stem-ML model (Kochanski et al., 2003) and the tone nucleus model (Zhang and Hirose, 2004). The Stem-ML model of Mandarin intonation takes as a starting point that all allophonic tonal variants are generated from lexical tonal templates and are distorted from these templates to a degree determined by their prosodic strengths (Kochanski et al., 2003), and the tone nucleus model for Mandarin tonal recognition attempts to extract parameters from the "tone nucleus" of each syllable, which represents underlying pitch targets corresponding to those in the citation form (Zhang and Hirose, 2004).

The use of acoustic parameters from preceding and following syllables and the privileged treatment of isolation forms in automatic tonal recognition systems raise cognitive issues about tonal representation. First, it is well known that *preceding* context both informs and biases tonal perception, but much less is known about the effect of *following* context. For instance, Wong and Diehl (2003) found that the fundamental frequency (f0) level of preceding syllables strongly biased the perception of level tones in Cantonese, and Chapter 3 shows that the f0 level of the preceding syllable also biases tonal identification in a forced choice task between the lowest level tone (˨) and fall (˨˩) in Cantonese; Huang and Holt (2009) also found that average f0 of preceding syllables affected the perception of the rise, a contour tone, in Mandarin (and see refs. within Wong and Diehl (2003) and Huang and Holt (2009) for many other studies of preceding context).

However, few tonal perception experiments have tested the effect of following context—in fact, few speech perception experiments at all have tested this (Lotto and Holt, 2006, p. 179); after all, speech perception unfolds in a forwards direction in time. The only one we know of for tonal perception is Gottfried and Suiter (1997), a small experiment with 5-6 listeners, which found that Mandarin listeners made fewer identification errors when the syllable to be identified was presented with a following syllable than without. Two studies, Xu (1994), and Francis et al. (2006), tested the effect of preceding and following context together, but not the effect of either context independently. Xu (1994) found that Mandarin listeners identified tones in "conflicting" contexts (coarticulated contexts in which the allophonic variant of the tone had a different f0 direction than in citation form) with much higher accuracy when presented in context with preceding and following syllables relative to when the flanking syllables were replaced with white noise. Francis et al. (2006) found that Cantonese listeners had sharper tonal identification boundaries between level tones when their sentential context was present (including both preceding and following context).

As far as we know, the only work *comparing* following and preceding context in tonal classification by human or machine is Levow (2005)'s automatic tonal recognition study of the effect of including preceding (left) vs. following (right) contextual parameters on automatic tonal recognition accuracy in Mandarin, which found that preceding context improved overall tonal recognition accuracy more than following context by around 5%. Levow (2005) writes that these results are "consistent with current linguistic theory which claims strong persistence or carryover effects in tonal coarticulation and only very weak anticipatory ones". Indeed, Xu (1997) found that left-to-right carryover coarticulation is stronger than right-to-left anticipatory coarticulation in the production of Mandarin bitones, and this has also been found in production studies of Cantonese bitones (Li et al., 2002; Flynn, 2003; Wong, 2006). However, we add to Levow (2005)'s interpretation of results that it is not clear that stronger carryover than anticipatory coarticulation implies that preceding acoustic context is uniformly more informative than following acoustic context for tonal classification: it depends on how one defines what information is associated with what linguistic unit, and how informativity is measured. For instance, if we define f0 values as being associated with a rise in a particular syllable if they show a positive velocity and are within a local temporal window defined with some bound, then the following acoustic context might share more mutual information (Cover and Thomas, 2006, p. 19-20) about the syllable of interest than the preceding, since the final ascent to peaks associated with rises is known to be delayed to the following syllable (Silverman and Pierrehumbert, 1990). Thus, it is of interest to compare tonal classification with preceding vs. following context available to the listener.

The second cognitive issue raised from automatic tonal recognition is whether the perspective of a special status for tones in isolation in Mandarin tonal recognition systems is useful for understanding the cognitive status of alternation in lexical tones in general. One reason that citation tones are taken to be underlying in the analysis of many Chinese tonal systems is that it is in isolation that the most contrasts are preserved, while tones merge in connected speech (Chen, 2000, 49). A natural question then is: are tones in isolation more perceptually distinct than tones in connected speech? There is a tension here: on the one hand, higher than 90% accuracy has been achieved in *speaker-independent* tonal recognition of isolated Mandarin tones using Hidden Markov Models (Yang et al., 1988) and automatic tonal recognition in connected speech is considered to be much more challenging than for isolated forms (Zhang and Hirose, 2004)—but there is also the intuition that "the most severe test of the phonological distinctiveness of tonal features would seem to be in the context-free condition of isolated monosyllabic words" (Abramson, 1972, 33). The Mandarin tonal inventory lacks level tone contrasts, but most tone languages have at least one level tone contrast (Maddieson, 1978). For tone languages with level contrasts, the lack of context in isolation would be expected to make perceptual normalization difficult, especially in speaker-independent tonal perception. Previous studies have compared tonal perception of monosyllables in connected speech and these monosyllables presented with neighboring syllables, e.g. Ma et al. (2005, 2006); Francis et al. (2006), but they have not studied tonal perception of monosyllables *in isolation* compared with tonal perception in connected speech.

To investigate the role of following context in tonal perception, and to compare the perception of isolated tones to tones extracted from connected speech, we performed a speaker-independent tonal perception experiment in Cantonese manipulating the context available to the listener and computationally modeled the listener's task in an acoustic space. We chose to study speaker-independent rather than speaker-dependent tonal perception because phonological categories traditionally are not conditioned on speaker identity, and we were interested in studying how much local acoustic contextual information and the potentially higher separability of tones in isolation would help listeners in this more challenging task. We chose Cantonese for the ease of finding a sample of speakers large enough for the experimental design and because the six tones of Cantonese comprise a good exemplar tone inventory in having both level (high level Tone 1, 55, ˥; mid level Tone 3, 33, ˧; low level Tone 6, 22, ˨), rising (high rising Tone 2, 35/25,˧˥/˨˥; low rising Tone 5 23/13, ˨˧/˩˧), and a falling tone (Tone 4, 21, ˨˩), cf. Figure 2.1 (Matthews and Yip, 1994).[1].

[1] Some descriptions also distinguish these tones from the shorter entering tones (high, mid, and low level) which occur in syllables with unreleased stop codas.

We asked: (i) *Can following context be as informative as preceding context, and is it informative in the same way?*, and (ii) *How does tonal perception compare for isolated tones vs. tones from connected speech?* We hypothesized that following context could improve tonal accuracy for contour tones, where there might be peak delay—when a f0 peak for a tone associated to a given syllable is realized following that syllable (Myers, 2003; Silverman and Pierrehumbert, 1990)—and that tonal perception in isolation would be worse for level tones than in connected speech, provided that contextual information was made available to the listener for perceptual normalization in connected speech. The rest of the paper is comprised of descriptions of the following: the speech materials used in the perception experiment and computational modeling (§2.2), the perception experiment (§2.3), and the computational modeling (§2.4). It concludes with a general discussion (§2.5).

2.2 Speech materials

2.2.1 Recordings

The stimuli were recorded by ten native Cantonese speakers, five of whose recordings were further processed for the rest of the study: these three males and two females were chosen to span overlapping pitch ranges over a wide overall range to be representative of the challenge of interspeaker variability in speaker-independent tonal recognition, also the task studied in the most recent automatic Cantonese tonal recognizers (Peng and Wang, 2005; Qian et al., 2007), cf. §2.2.3, Table 2.1 for range information. One of the speakers was born and raised in Macau and recorded in the phonetics lab sound-attenuated booth at University of California, Los Angeles. The other four were born and raised in Hong Kong and recorded in the phonetics lab sound-attenuated booth at the City University of Hong Kong. The speakers were recruited from the local university student population and received cash compensation. All were recorded digitally at 22,050 Hz/16 bits with PCQuirerX (Scicon R&D, Inc.) or at 44.1kHz/16 bits with a digital recorder.

The stimuli were created from: (i) the citation/isolation form of all six tones over /wai/, one of the few sonorant syllables over which all tones are familiar real words, and from connected speech, (ii) the tritone ⟨ *wai*˧, {*wai*˥, ˧˥, ˧, ˨˩, ˨˧, ˨} , *mat*˧ ⟩ (*wai*^{33} *wai mat*^{3}) extracted from sentences of the form: *lei*$^{25/35}$ *yiu*^{33} *wai*^{33} *wai mat*^{3} *deng/geng*^{33} '*you want wai-wai to clean the lamp/mirror*' with the target, the second /wai/, ranging over all six tones Tone 55 to Tone 22. A sonorant syllable was chosen so that the f0 contours could be analyzed over the entire syllable. The lexical meanings of the orthographic characters used to label the tones Tone 55-Tone 22 were, respectively, 'power', 'appoint', 'fear', 'surround', 'great', and 'stomach', and speakers were asked to treat /wai wai/ as a (nonce) proper name. The orthographic characters, which were the same for both eliciting recordings and the perception experiment, were chosen to be the most familiar ones for each tone by a native speaker. Each speaker recorded 5 fluent repetitions each of the citation tones, and of sentences containing all 36 bitone combinations over /wai wai/. A Cantonese native speaker checked that none of the speakers had tonal mergers and that the speakers uttered the correct tones.

From the recordings of isolated tones, we selected 3 repetitions per speaker per tone, for a total of 90 utterances. From the recordings of connected speech, we chose three repetitions of each Tone 33-Tx bitone for the stimulus set, for a total of 90 tritones, 18 from each speaker, 3 distinct repetitions per speaker per Tone 33-Tx bitone. In pilot studies, the stimuli consisted of all 36 licit bitones, rather than a subset of tritones, but subjects performed poorly and were confused by the many different stimulus types. Thus, the *wai wai* bitones were restricted to having an initial Tone 33 tone, since Tone 33 provides a kind of neutral context as it is the mid-range level tone in Cantonese, and the following *mat*^{3} (also mid-range level) was included as well to form a tritone.

2.2.2 Resynthesis

All stimuli were resampled to 22kHz; tritones were extracted using a rectangular window, and RMS amplitude was rescaled to 75 dB in Praat (Boersma and Weenink, 2010). All syllable durations were resynthesized using PSOLA implemented in Praat to be have a target duration of 241±43 ms (SD), the grand mean of the syllable durations, for a total duration of 740 ms for the tritone; the isolation syllables were resynthesized to their grand mean of 512±108 ms (SD).[2] Amplitude normalization was performed as a control to standardize amplitude across the recordings from different speakers and tones, and duration normalization was performed to create conditions for the listener similar to the time-normalized parameter extraction in Peng and Wang (2005); Qian et al. (2007) and many other automatic tonal recognizers, where f0 values are extracted uniformly over the syllable.

Besides the isolation condition ISO, four additional conditions of the manipulated variable CONTEXT were nested within the tritones and created by restricting the stimuli to bitones and monotones,[3] listed below with the target syllable indicated in boldface: the tritone condition TRI (*wai*^{33} ***wai*** *mat*^{3}), the bitone conditions with the pre-target syllable providing preceding context, PRE (*wai*^{33}

[2] The PSOLA algorithm resynthesis added about 18 ms over the target duration over the course of the tritone.

[3] In this paper, *monotone* always refers to a sequence of a single tone, just as bitone refers to a sequence of two tones. It does not refer to a flat pitch.

wai) and with the post-target syllable providing following context, POST (***wai*** *mat*[3]), and the monotone condition MONO (***wai***).

2.2.3 Acoustic analysis of resynthesized speech materials

We performed an acoustic analysis of the resynthesized speech materials based on extracted f0 tracks. The f0 values were extracted using RAPT (Talkin, 1995), a commonly used f0 detection algorithm, used in Qian et al. (2007)'s Cantonese supratone model and other tone recognizers. Speaker-specific pitch floors and ceilings were set following the preprocessing procedures in De Looze and Rauzy (2009); Evanini and Lai (2010), with speaker pitch range estimated using the 35th/65th quantiles for the isolation condition, and with the 1st and 99th quantiles minus or plus 30% of the range, respectively, for the tritones. The majority of the f0 values for the tritones were in the mid range since each stimulus consisted of two mid-level tones, yielding a center-heavy distribution of f0 values; thus, we could not use the 35th quantile-based calculation from De Looze and Rauzy (2009), since the center-heavy distribution resulted in large compression in the range estimation. Otherwise, the default RAPT parameter settings, including a 10ms frame shift, were used. The first and last frames were excluded because they were often assigned f0 values creating large jumps to f0 of the adjacent frame. There were a total of 46 f0 values in the ISO condition (Figs. 2.5 and 2.6). There was a total of 69 f0 values for the TRITONE condition (Fig. 2.1); from these f0 values, subsets of values were extracted for the other context conditions, e.g. the middle 23 values for the MONO condition and the first 46 for the PRE condition. Unvoiced frames were assigned f0 values using linear interpolation. The f0 values were also log-transformed and then standardized as z-scores using speaker-specific means and standard deviations. The calculated raw and transformed f0 range of the stimuli for each speaker is given in Table 2.1. The parameterized stimuli were used as data for computational modeling, described in §2.4.

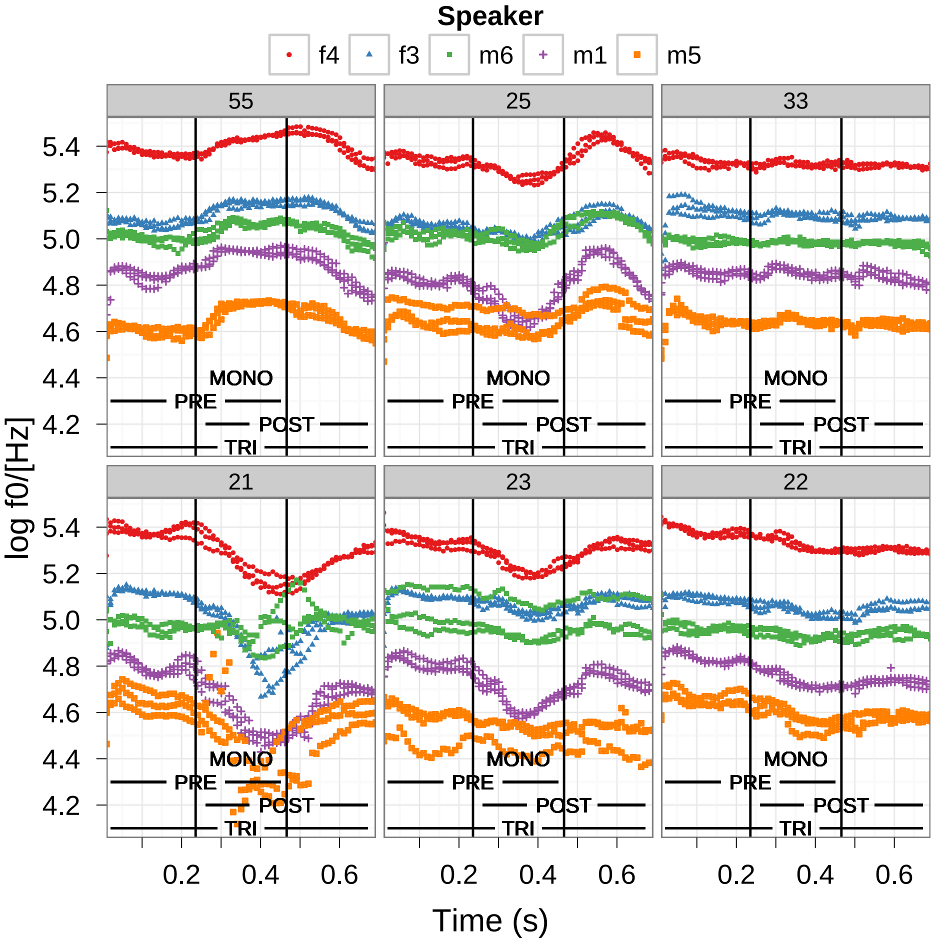


Figure 2.1: f0 contours of the tritone stimuli extracted for each of the 5 speakers using speaker-specific pitch floors and celings in RAPT. The frame shift is 10ms.

Speaker	f0 (Hz)	log f0	z-score
f4	[165.89,241.00]	[5.11,5.48]	[-3.35,2.35]
f3	[106.42,179.47]	[4.67,5.19]	[-5.78,1.83]
m6	[125.88,176.36]	[4.84,5.15]	[-2.91,2.94]
m1	[83.87,145.92]	[4.43,4.98]	[-3.48,1.97]
m5	[61.44,140.20]	[4.12,4.94]	[-5.08,3.60]

Table 2.1: Speaker-specific f0 range in speech materials, measured in Hz, after log-transformation, and after standardization of log f0, with respect to speaker means and standard deviations. The speakers are ordered from highest to lowest maximum f0.

2.3 Tonal perception experiment

Using the speech materials described in the preceding section, §2.2, we performed a human tonal perception experiment with Cantonese native speakers.

2.3.1 Methods

2.3.1.1 Participants

There were 18 male (age 20.9±1.9 years) and 18 female (age 21.9±1.9 years) native Cantonese speakers who participated. They were recruited from the local university student population at the City University of Hong Kong and at the University of California, Los Angeles and received cash compensation. All but one of the subjects (born/raised in Guangzhou, China) was born and/or raised in Hong Kong, China. Of the 8 participants tested in Los Angeles, all spoke Cantonese on a daily basis and had been in the United States for 2 to 5 years.

2.3.1.2 Procedure

Participants were told that the stimuli were extracted from sentences *lei*$^{25/35}$ *yiu*^{33} *wai*^{33} *wai mat*^{3} *geng*^{33} 'You want NAME to clean the mirror.' The stimuli were blocked by context (mono, pre, post, tri, iso) so that participants wouldn't be confused about which tone they were to identify, and the block order was pseudorandomized such that the context factor levels were roughly uniformly distributed over the five blocks presented. Stimuli order within blocks was randomized. Participants received a short break between blocks, and for each block, participants were given a sheet of paper with orthographic characters which showed what stimulus was being played, and what word they were to identify for each context: __ (iso), __ (mono), *wai*^{33} __ (pre), __ *mat*^{33} (post), and *wai*^{33} __ *mat*^{33} (tri).

The task of the participants was to lexically identify the target syllable in each stimulus by a keyboard press of one of six keys labeled with the characters for the minimal tone set over *wai*. Participants were asked to respond as quickly and accurately as possible and told they would be timed. Their responses and reaction times, measured from the onset of the stimuli, were recorded.

2.3.1.3 Data analysis

Statistical analysis was performed in R (R Development Core Team, 2010), and the ggplot2 package was used for creating graphics (Wickham, 2009). Tonal identification accuracy was analyzed using mixed effects linear regression implemented by the lme4 package (Bates and Maechler, 2010), a statistical method that has become common in language research (Baayen, 2008). The interest in this study, as in most psychological studies, was in generalizing beyond the sample of listeners and the sample of speakers from which the stimuli were drawn—in this case, to native Cantonese speakers.

Mixed effects models allow the inclusion of both the subject and the speaker as crossed (independent rather than nested) random effects (Baayen et al., 2008). In comparison, repeated measures ANOVA (RM-ANOVA) analyses allow only one such random effect in a model, and in the case of multiple, independent random effects as in this study, a standard practice is combinining results from separate RM-ANOVAs for each random effect (Clark, 1973). Using mixed-effects models allows simultaneous generalization to other participants and speakers, as both random effects are included in a single model (Quené and van den Bergh, 2008).

Forward model selection was used to test the partial effects of CONTEXT on tonal identification accuracy, and successive nested models were compared using likelihood ratio tests. Because the model likelihood (the probability of the data given the estimated model parameters) can always be increased by increasing the number of model parameters, χ^2 tests were used to test for significant improvement in fit to the data while penalizing for model complexity (Baayen, 2008, p. 253), as differences in deviance ($-2\log(\mathit{likelihood})$) between nested models fit to the same data by maximum likelihood approximately follow a χ^2 distribution in the large-sample limit.

For multiple comparisons on the mixed effects models, the multcomp package (Hothorn et al., 2008) was used, with Tukey tests for all pairwise comparisons of CONTEXT and Bonferroni tests for selected pairwise comparisons.

2.3.2 Results

In this section, we report on the perception experiment results focusing on: (i) comparing the two bitone conditions, PRE and POST in the context of the MONO and TRI conditions and directly to one another (§2.3.2.1), and (ii) comparing the

ISO condition with two other conditions from connected speech—the other monosyllable condition, MONO and the condition with the maximal local contextual information in the experiment, the tritone condition TRI (§2.3.2.2). Within each of these subsections, we report on results aggregated across tones followed by results for individual tones.

2.3.2.1 Bitone context conditions

Results aggregated across tones Overall tonal identification accuracy for connected speech stimuli was significantly affected by CONTEXT and was significantly higher as more syllables were available to the listener: ordered from lowest to highest accuracy, the CONTEXT conditions were MONO, POST, PRE, TRI (Figure 2.2). Accuracy was significantly higher for the PRE than the POST condition. The significant effect of CONTEXT was established with model comparison—between a linear mixed effects model for tonal identification accuracy with random intercepts by-subject and by-(stimuli) speaker and one with the addition of a fixed effect for CONTEXT (MONO, PRE, POST, TRI)—which supported the inclusion of CONTEXT ($\chi^2(3) = 184.92$, $p < 2.2 \times 10^{-16}$). The model with CONTEXT and associated multiple comparisons are shown in Table 2.2, for which every pairwise comparison showed a significant difference.

Results for individual tones Model comparison supported the inclusion of CONTEXT in models of tonal identification accuracy for every individual tone except Tone 23 ($\chi^2(3)$ = 204.71, 260.91, 189.07, 29.45, 6.01, 20.95, for Tone 55-Tone 22, respectively; $p < 2.2 \times 10^{-16}$ for Tone 55-Tone 33, 1.8×10^{-6} for Tone 21, 0.11 for Tone 23, 1.1×10^{-4} for Tone 22). Multiple comparisons between CONTEXT conditions for individual tones are summarized in Table 2.3 and Fig. 2.3. The

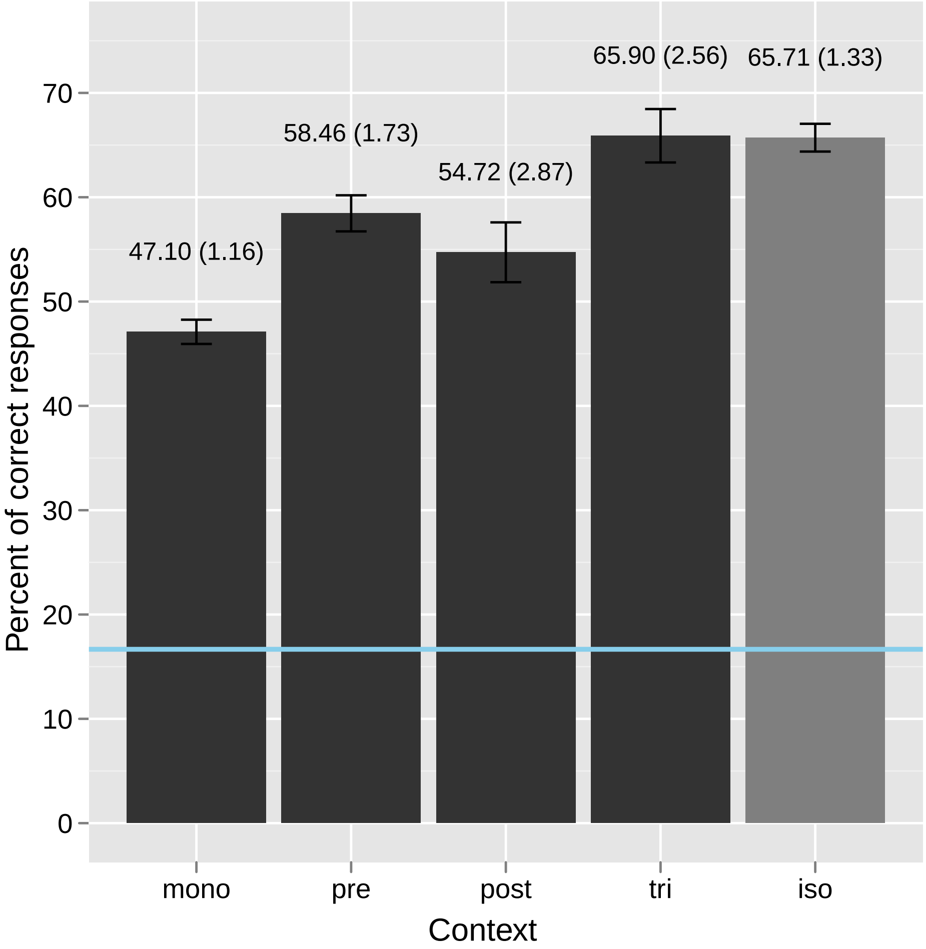


Figure 2.2: Comparison of tonal identification accuracy for different local acoustic context conditions. For all contexts, Bonferroni-corrected t-tests of by-subject tonal ID accuracy against the at-chance level (indicated by the horizontal line: 1/6, 17%) showed that performance for each condition was signficantly above chance. Error bars show ±1SE.

	Coef β	SE(β)	z	p
MONO - PRE	-11.36	1.32	-8.62	<0.0001
MONO - POST	-7.62	1.32	-5.79	<0.0001
POST - PRE	-3.74	1.32	-2.84	<0.024
MONO - TRI	-18.80	1.32	-14.27	<0.0001
POST - TRI	-11.17	1.32	-8.48	<0.0001
PRE - TRI	-7.44	1.32	-5.65	<0.0001

Table 2.2: Multiple comparisons of context contrasts for tonal identification accuracy in stimuli from connected speech, calculated using Tukey tests on a linear mixed effects model with random intercepts by-subject and by-speaker and CONTEXT as a fixed effect. A negative regression coefficient β indicates a higher model estimate of tonal identification accuracy in the right member of the comparison.

Tone 23 rise was not significantly affected by context in any comparison, and confusion matrices conditioned on CONTEXT showed similar confusion patterns for all conditions in connected speech (Fig. 2.3, Table 2.5); Tone 23 was confused with the other rise Tone 25, and also the mid to low level tones, Tone 33 and Tone 22, in particular.

Among the level tones, in comparisons between contexts, Tone 55 and Tone 33 were identified with significantly higher accuracy in conditions where the pre-target syllable was present, but Tone 22 was identified significantly more accurately when the post-target syllable was present. Tone 55 and Tone 33 accuracies were significantly higher in the PRE than the POST condition, while the opposite was true for Tone 22 accuracy, and between the MONO and POST conditions, Tone 55 accuracy was actually significantly lower for the POST condition, but was significantly higher for the PRE condition than the MONO condition; Tone 22 accuracy was not significantly different between the MONO and PRE conditions, but showed significant improvement in the POST condition relative to the MONO condition. From the bitones to the TRI condition, there was significant improvement for Tone 55 and Tone 33 accuracy from the POST condition, but no significant difference for Tone 33 accuracy between the PRE and TRI conditions, and actually a significant decrease in Tone 55 accuracy in the TRI condition relative to the PRE condition.

From the inspection of confusion matrices conditioned on CONTEXT (Fig. 2.3, Table 2.5), we found that the level tone stimuli in the PRE condition were systematically confused mostly with higher level tones, while they were systematically confused mostly with lower level tones in the POST condition. Also, the confusability in the PRE condition of level tones with higher level tones was lower than the confusability in the POST condition of level tones with lower level tones. For instance, for Tone 33, the most confusable tone in the PRE condition was Tone 55 (17.04%), while the most confusable tone in the POST condition was Tone 22 (36.11%).

As for the contour tones, the Tone 25 rise and Tone 21 fall were identified with significantly higher accuracy in contexts with the post-target syllable available in context comparisons, and Tone 25 was also identified with significantly higher accuracy with the pre-target syllable available in most context comparisons. Between the two bitone conditions, Tone 25 was identified significantly more accurately in the POST condition; however, there was significant improvement from the MONO condition to both bitone conditions, and also from both bitone conditions to the TRI condition. There was no significant difference in Tone 21 accuracy between the two bitone conditions, but from the MONO condition to the bitone conditions, Tone 21 accuracy was significantly higher in the POST condition but not significantly different in the PRE condition; from the bitone conditions to the TRI condition, there was no sigificant difference in Tone 21 accuracy compared to the POST condition, but there was significantly higher accuracy compared to the PRE condition.

Analysis of the confusion matrices showed that the presence of the post-target syllable reduced confusion of Tone 25 with Tone 33 and confusion of Tone 21 with Tone 22. Both the presence of the pre-target syllable and the presence of the post-target syllable reduced confusability of Tone 25 with Tone 23 from the monosyllabic condition (67.41%) to the bitone conditions (49.63%, 48.33% for PRE and POST, respectively), and the presence of both the pre- and post-target syllables together in the TRI condition further dropped confusability with Tone 23 to 25.74%.

2.3.2.2 The isolation context

Results aggregated across tones Model comparison supported the inclusion of CONTEXT in modeling overall tonal identification accuracy for the MONO, TRI and ISO conditions ($\chi^2(2) = 226.53$, $p < 2.2 \times 10^{-16}$). Multiple comparisons with Bonferroni adjustments showed that accuracy was significantly higher in isolation than for monosyllables extracted from connected speech ($p < 2e^{-16}$), but there was no significant difference between accuracy in isolation and for the TRI condition ($p = 1$).

Results for individual tones Tonal identification accuracy was significantly higher for isolated monosyllables than monosyllables extracted from connected speech for every tone except Tone 55, for which accuracy was surprisingly significantly higher for the MONO condition (Table 2.4). The confusion matrices (Fig. 2.3, Table 2.5) showed that there was more confusability of Tone 55 with Tone 33 in isolation than in the MONO condition. Between the ISO and TRI conditions, accuracy was significantly higher for Tone 55, Tone 25, and Tone 33 for the tritones, but accuracy was significantly lower for Tone 23 and Tone 22 in isolation, and there was no significant difference in accuracy for Tone 21. The level tones Tone 55 and Tone 33 were mostly confused with lower level tones, Tone 33 and Tone 22, respectively, in isolation. The lowest level tone Tone 22 was confused mostly with Tone 33 and also Tone 21 in isolation, but with the rises Tone 25 and Tone 23 as well as Tone 33 and Tone 21 for the TRI condition.

Fig. 2.3 highlights that listeners were strikingly more accurate on Tone 23 in the ISO condition than in any of the connected speech conditions, in which CONTEXT had no significant effect on accuracy. The higher accuracy for Tone 23 in isolation does not seem to be due to a strong response bias for Tone 23 in ISO as the response frequency for Tone 23 in ISO was 22%, higher than in TRI (17%), but similar to the response frequency in PRE and POST (21%), and lower than the response frequency in MONO, 25%. Nor does it seem to be due to a lexical bias. As a rough estimate of lexical frequencies of the six orthographic characters used in the identification task, we used the frequencies of the Mandarin cognates, [wei], in the character frequency list of Modern Chinese from Da (2004). Counts from that text corpus indicated the following relative frequency percentiles, from the most to least frequent character used to represent the tones: Tone 25 (26), Tone 21 (21), Tone 55 (20), Tone 23 (9), Tone 22 (3), Tone 33 (3); thus the estimated lexical frequency of the character used for Tone 23 was relatively low.

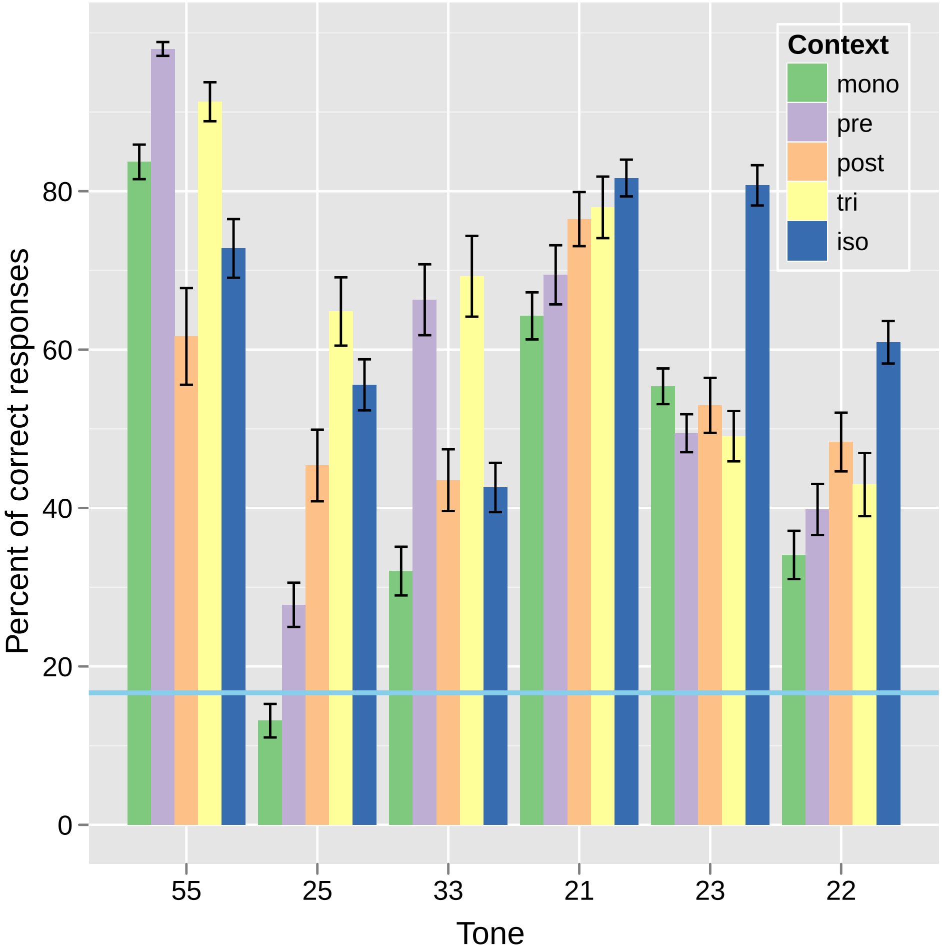


Figure 2.3: Comparison of tonal identification accuracy for different local acoustic context conditions, grouped by individual tone. For every context, performance was well-above chance (the horizontal line shows identification accuracy for at-chance performance (1/6, 17%)), and the error bars show ±1SE.

	Tone 55	Tone 25	Tone 33	Tone 21	Tone 23	Tone 22
MONO - PRE	PRE, $p < 0.001$	PRE, $p = 1e^{-5}$	PRE, $p < 1e^{-4}$	n.s., $p = 0.27$	n.s., $p = 0.20$	n.s., $p = 0.27$
MONO - POST	MONO, $p < 0.001$	POST, $p < 1e^{-5}$	POST, $p = 7.4e^{-4}$	POST, $p < 0.001$	n.s., $p = 0.85$	POST, $p < 0.001$
POST - PRE	PRE, $p < 0.001$	POST, $p < 1e^{-5}$	PRE, $p < 1e^{-4}$	(POST), $p = 0.066$	n.s., $p = 0.64$	POST, $p = 0.037$
MONO - TRI	TRI, $p = 0.013$	TRI, $p < 1e^{-5}$	TRI, $p < 1e^{-4}$	TRI, $p < 0.001$	n.s., $p = 0.15$	TRI, $p = 0.026$
PRE - TRI	PRE, $p = 0.039$	TRI, $p < 1e^{-5}$	n.s., $p = 0.75$	TRI, $p = 0.015$	n.s., $p = 0.99$	n.s., $p = 0.75$
POST - TRI	TRI, $p < 0.001$	TRI, $p < 1e^{-5}$	TRI, $p < 1e^{-4}$	n.s., $p = 0.95$	n.s., $p = 0.57$	n.s., $p = 0.33$

Table 2.3: Summary of multiple comparisons by individual tone. The condition that yielded significantly more accurate tonal identification accuracy, or n.s. for a nonsignificant difference, is listed along with the estimated p-value.

	Tone 55	Tone 25	Tone 33	Tone 21	Tone 23	Tone 22
MONO - ISO	MONO, $p = 3.5e^{-5}$	ISO, $p < 2e^{-16}$	ISO, $p = 1.8e^{-3}$	ISO, $p = 2.4e^{-10}$	ISO, $p < 2e^{-16}$	ISO, $p < 2e^{-16}$
TRI - ISO	TRI, $p = 7.0e^{-13}$	TRI, $p = 2.5e^{-3}$	TRI, $p < 2e^{-16}$	n.s., $p = 0.34$	ISO, $p < 2e^{-16}$	ISO, $p = 1.2e^{-8}$

Table 2.4: Summary of multiple comparisons by individual tone for comparisons with isolation condition. The condition that yielded significantly more accurate tonal identification accuracy, or n.s. for a nonsignificant difference is listed along with the p-value estimated.

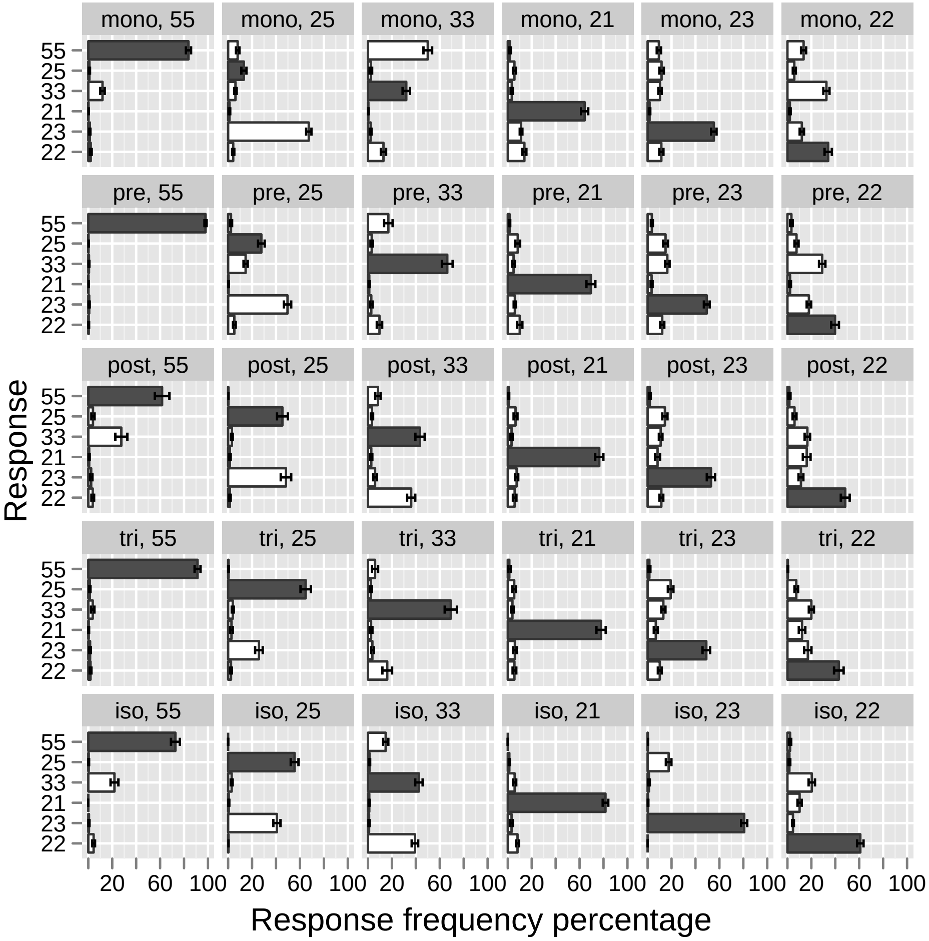


Figure 2.4: Visualization of confusion matrix with CONTEXT (row) and TONE (column) conditions completely crossed. Percentage of response frequency for a given tone is given aggregated over subjects and stimulus speakers. Error bars indicate ±1SE estimated by-subjects.

Actual	**Response**					
	Tone 55	Tone 25	Tone 33	Tone 21	Tone 23	Tone 22
Tone 55						
Mono	83.70	0.93	11.85	0.19	1.30	2.04
Pre	97.96	0.19	0.56	0.19	0.74	0.37
Post	61.67	3.89	27.59	0.74	2.41	3.70
Tri	91.30	1.30	3.70	0.37	1.48	1.85
Iso	72.78	0.37	21.85	0.00	0.56	4.44
Tone 25						
Mono	7.96	13.15	6.11	1.11	67.41	4.26
Pre	2.41	27.78	14.63	0.37	49.63	5.19
Post	0.19	45.37	3.15	1.48	48.33	1.48
Tri	0.37	64.81	3.89	2.78	25.74	2.41
Iso	0.00	55.56	2.96	0.56	40.74	0.19
Tone 33						
Mono	50.00	2.41	32.04	0.37	2.22	12.96
Pre	17.04	3.15	66.30	1.11	2.78	9.63
Post	8.33	3.33	43.52	2.78	5.93	36.11
Tri	5.93	2.41	69.26	2.59	3.70	16.11
Iso	14.81	1.30	42.59	1.11	0.93	39.26
Tone 21						
Mono	1.85	5.56	3.33	64.26	11.11	13.89
Pre	1.48	8.33	4.81	69.44	5.93	10.00
Post	0.74	6.48	3.15	76.48	7.41	5.74
Tri	1.30	5.37	3.89	77.96	5.93	5.56

Iso	0.00	1.30	5.74	81.67	3.15	8.15
Tone 23						
Mono	9.44	11.67	10.37	1.67	55.37	11.48
Pre	3.52	15.00	16.48	3.33	49.44	12.22
Post	1.85	14.44	10.93	8.33	52.96	11.48
Tri	1.48	19.26	13.15	6.85	49.07	10.19
Iso	0.19	17.59	1.11	0.37	80.74	0.00
Tone 22						
Mono	13.52	5.74	32.59	2.04	12.04	34.07
Pre	3.33	7.59	29.07	2.22	17.96	39.81
Post	1.67	5.93	16.67	16.11	11.30	48.33
Tri	0.37	7.41	20.00	12.22	17.04	42.96
Iso	2.22	1.67	20.37	10.19	4.63	60.93

Table 2.5: Confusion matrix conditioned on tones by context.

2.3.3 Discussion

We first discuss the bitone context comparison and then the isolation context comparisons. Results from the perception experiment indicated that the post-target syllable could be as informative as the pre-target syllable for tonal identification of the target syllable, and that the pre- and post-target syllables were informative in complementary ways. Based on comparisons of tonal identification accuracy between the MONO, PRE, POST, and TRI conditions, the pre-target syllable was crucial for maximizing Tone 55 identification accuracy under the experimental conditions, and the post-target syllable for Tone 21 and Tone 22 identification accuracy. For Tone 25 and Tone 33, both the pre- and post-target syllables were informative for the listener, but in head-to-head comparisons between the two bitone comparisons, identification accuracy for Tone 25 was higher with the pre-target syllable, while Tone 33 accuracy was higher with the post-target syllable. Tone 23 was the only tone for which the two bitone conditions showed no differences for any identification accuracy comparisons, but Tone 23 showed no effect of context at all among the connected speech conditions.

The advantage of the post- over the pre-target context for identifying the contour tones Tone 25 and Tone 21 is likely driven by the peak delay in the f0 contours for Tone 25 and Tone 23 into the post-target syllable relative to the pre-target and target syllables (Wong, 2006) and the presence of the post-target rise in Tone 21 but not Tone 22 stimuli. The peak delay for Tone 25 and the post-target rise for Tone 21 are represented in the mean log-transformed f0 contours for Tone 25 and Tone 23 in Fig. 2.5 and in the comparison of the Tone 21 and Tone 22 f0 contours in Fig. 2.1. Without cues available that were delayed to the post-target syllable, even the larger rise in the tonal inventory, Tone 25, was frequently perceived as a level tone, and the only fall (in isolation), Tone 21, was frequently perceived as the lowest level tone, Tone 22.

A potential caveat for the Tone 25 results is that a grammatical factor could be in play. In specific morphological contexts such as for vocatives for kinship terms and intensification for adjectives, reduplicated expressions in Cantonese such as the bitone $wai^{33}wai^{33}$ are the target of a process of tone change in which the second tone may change into a rise identical or similar to Tone 25 (Chen, 2000; Matthews and Yip, 1994; Yu, 2007, 2009), so that there is a possibility that listeners would have considered a Tone 25 target stimulus in the PRE condition to be the changed form of a Tone 33 and identified the target as an underlying form Tone 33. While we might argue that the morphological/semantic context in the experiment was not one where tone change is active, we also confirmed that the listener responses for Tone 25 were not consistent with active tone change.[4]

The advantage of the pre- over the post-target context for identifying Tone 55 and Tone 33 was systematically due to confusion in the POST condition with lower level tones, Tone 33 and Tone 22, respectively. Given that for both Tone 55 and Tone 33, the f0 contour fell from the target syllable through the post-target syllable, especially after Tone 55 target syllables (see right panel in Fig. 2.6), the listener response to confuse the target tone with a lower level tone in the POST condition seems unintuitive at first glance.

Some reasons for the listener confusion with lower level tones in the POST

[4] First, if the confusability of Tone 25 with Tone 33 in the PRE condition was driven by some listeners responding with the underlying form Tone 33, then we would expect some listeners with a majority of Tone 33 responses. However, the highest proportion of Tone 33 responses for Tone 25 in the PRE condition for any listener was 40%; the mean proportion was $14.63\% \pm 10.55$ (SD) and only 4 of 36 listeners had no Tone 33 responses at all. But what if the listeners didn't perceive a majority of Tone 25 stimuli in the PRE condition as Tone 25 surface forms? We confirmed that the proportion of Tone 33 responses in the PRE condition was not the highest for stimuli from the speaker with the most citation-form like Tone 25 forms in the PRE condition, m1 (Fig.2.1). In fact, m1 Tone 25 stimuli in the PRE condition yielded the fewest Tone 33 responses, while m5 stimuli yielded the most, suggesting that the Tone 33 confusability was due to acoustic separability between Tone 25 and Tone 33 rather than the effect of tone change.

condition could be that: (i) listeners are inherently poor at integrating contextual information after the target to be identified in perceptual normalization; for instance, perhaps listeners were influenced by an expectation of declination (Vance, 1976; Flynn, 2003; Wong and Diehl, 2003; Ma et al., 2005), and overcompensated for it (ii) carryover coarticulation is stronger than anticipatory coarticulation (Flynn, 2003; Wong, 2006), so the post-target syllable did not provide as much target-extrinsic information as the pre-target syllable for perceptual normalization, (iii) listeners may have interpreted the stimulus-final fall in the post-target syllable as a strong downtrend due to sentence-level prosody, such as a declarative-final fall (Vance, 1976; Zee, 1998; Wong et al., 2005), not carryover coarticulation from the target syllable, even though we specified that the stimuli were extracted from sentence-medial position.

We restrict reasons (i) and (ii) to perceptual normalization since, given the informativity of the POST context for Tone 25, it seems that listeners can use contextual information after the target and that the post-target syllable can be more informative than the pre-target syllable for a given tone—but perhaps integration of context for level tones is different. While listeners' expectations about declination and how this affects Cantonese tonal perception has been investigated in Wong and Diehl (2003); Ma et al. (2005, 2006), only preceding context has been studied, and to our knowledge, it is not known how well listeners integrate following context.

The effect of an expectation of declination or fall is supported by results for the lowest level tone Tone 22. These did not follow the pattern of results that level tones were more accurately identified when the pre-target syllable was available than when the post-target syllable was available because Tone 22 was more confusable with Tone 33 in the PRE condition—more confusable than Tone

22 was with Tone 21 in the POST condition, the only lower tone than Tone 22. Tone 22 results are consistent with the listener's expectation of a downtrend in the final syllable played, whether that be the target syllable, as in the PRE condition, or the post-target syllable, as in the POST condition. Also, for Tone 55 there is no higher level tone available as a competitor, and for Tone 22 there is no lower level tone available as a competitor: we conjecture this is why the biggest differences in accuracy between the two bitone conditions occurred for Tone 55 and Tone 22, and in opposite directions. Overall, though both the PRE and POST conditions may have been effected by expectations about downtrends, the PRE condition produced higher identification accuracies than the POST condition for the level tones, while the POST condition produced higher accuracies for the contour tones Tone 25 and Tone 21. We discuss results for Tone 23 after discussing results for tones in isolation.

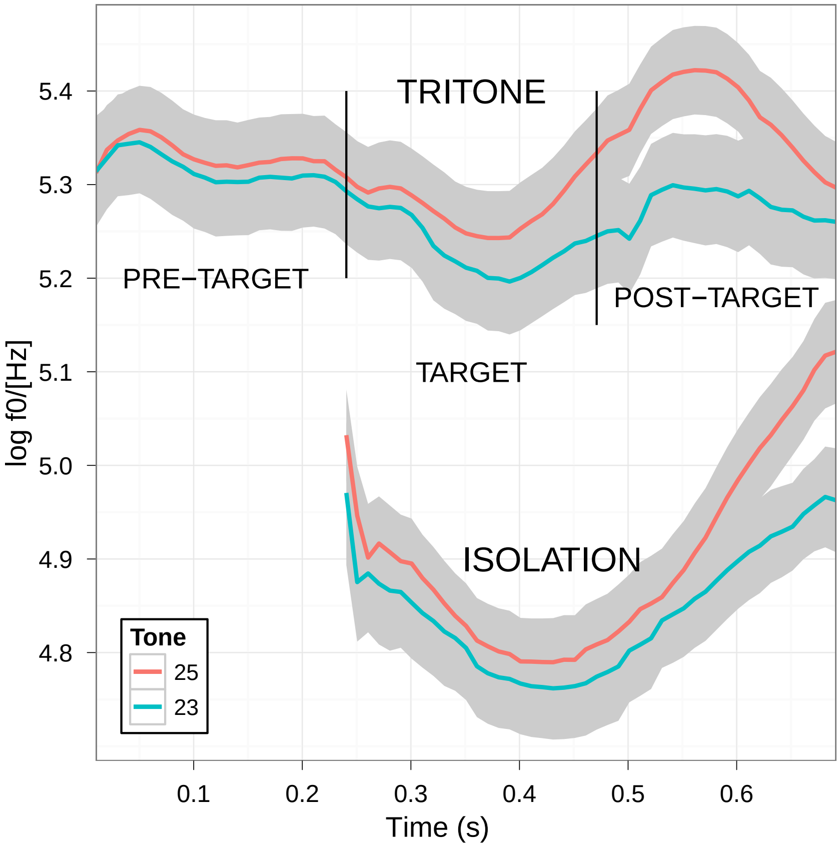


Figure 2.5: Mean log-transformed f0 contours for the two rises, Tone 25 and Tone 23, for the tritone and isolation stimuli. The tritone f0 contours have been vertically shifted by 0.35 to offset them from the isolation stimuli, whose onsets have been shifted to coincide with the onsets of the target syllable in the tritone stimuli. The ribbons show ±1SE, and the vertical dividing lines show the onsets of the 2nd and 3rd syllables in the tritone.

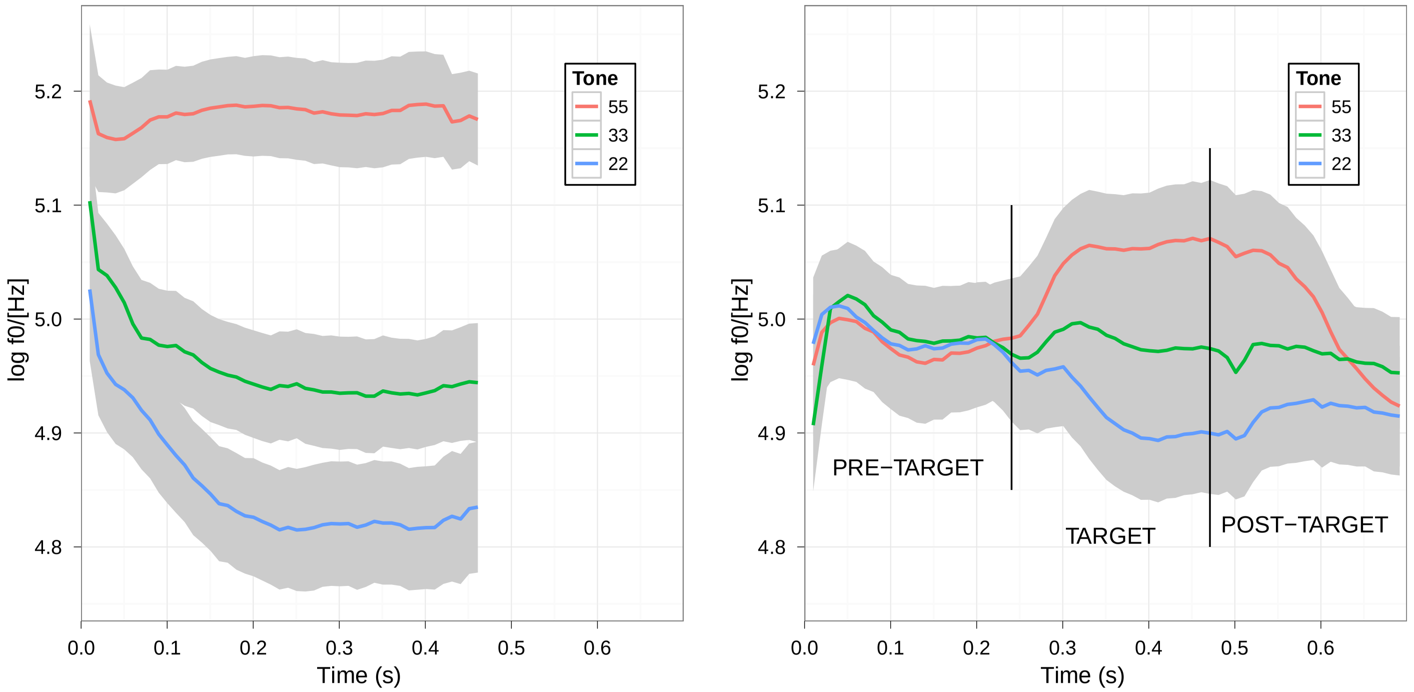


Figure 2.6: Mean log-transformed f0 contours for the three level tones, Tone 55, Tone 33, and Tone 22, for the tritone (l) and isolation stimuli (r).The ribbons show ±1SE, and the vertical dividing lines show the onsets of the 2nd and 3rd syllables in the tritone. Both figures use the same x- and y-axis scales.

Turning to the perception of the isolated tones, it was not surprising that overall, tonal identification accuracy was higher between the two monosyllabic conditions for the ISO context than the MONO context extracted from connected speech. First, the isolation stimuli were about twice as long in duration, as we used the grand means of syllable durations in isolation and in connected speech separately in fixing the stimuli duration. Second, Figs. 2.5 and 2.6 suggest that f0 contours of the rises Tone 25 and Tone 23 and of the level tones Tone 55, Tone 33, and Tone 22 were more separable in isolation than in connected speech, even aggregated across the five speakers—in speaker-independent tonal recognition: the f0 contours within the set of rises and the set of level tones were more separated, and the f0 contours across the sets were more separated. Both the MONO and ISO conditions lacked syllable-extrinsic information for perceptual normalization, so the higher separability of the syllable-instrinsic f0 contours of tones in isolation explains all the results for individual tones showing higher identification accuracy in isolation.

This makes the result of higher identification accuracy in the MONO condition for Tone 55, the level tone in the relatively uncrowded high f0 range, all the more shocking: why should Tone 55 in isolation be more confusable with Tone 33 than in extracted monosyllables from connected speech? We suspect that carryover coarticulation at the onset of the target syllable—the rise into a level high f0 contour (see the Tone 55 panel in Fig. 2.1)—provided "syllable-extrinsic" information that was realized in the target syllable, in the sense that that initial rise within the target syllable was conditioned on the previous mid-level Tone 33. While a fall at the target onset following a Tone 33 might indicate any one of the tones in the lower region of the pitch range, a rise to the high region of the pitch range singles out Tone 55.

It was also surprising that tonal identification accuracy was similar between the isolation context—with no syllable-extrinsic information—and the tritone context—with pre- and post-target syllables available to the listener, but this seems reasonable upon examination of results split by pitch register. As noted in §2.1, Wong and Diehl (2003) established that preceding context dramatically biases the perception of Cantonese level tones, and Francis et al. (2006) found that the presence of surrounding sentential context sharpens the identification boundary between the level tones. In support of this, Tone 55 and especially Tone 33 were identified with higher accuracy in the tritone condition than in the isolation condition, in which they were confused with other level tones, as discussed earlier. However, Tone 22, the other level tone, was identified overall with higher accuracy in the isolation condition. This was due to confusability of Tone 22 with not just the next highest level tone Tone 33, but also other low-range tones, Tone 21 and Tone 23: level and contour tones were more confusable with each other in connected speech than in isolation.

The higher level tones were identified more accurately in the TRI condition, while the lowest level tone was identified more accurately in isolation; similarly, the rise spanning a larger and higher pitch range, Tone 25, was identified more accurately in the TRI condition, while the smaller, lower rise, Tone 23, was identified more accurately in isolation. Like for Tone 22, the higher accuracy for Tone 23 in isolation was due to the higher confusability of contour tones with level tones in connected speech, cf. Fig. 2.5 whereas in isolation, the rises were only confused with each other (Fig. 2.4, 2.5).

It is surprising that Tone 25 was identified more accurately in isolation than in the TRI condition: unlike Tone 23, Tone 25 showed only negligible confusability with level tones in connected speech and was more confusable with Tone 23 than in the TRI condition. We conjecture that in an effectively binary decision between the two rises, listeners performed better with connected speech from both the availability of the pre-target syllable for perceptual normalization, as in Huang and Holt (2009), and the availability of the post-target syllable due to peak delay: the post-target syllable helped listeners rule out level tones as candidates, and the pre-target syllable helped them gauge the size of the rise.

Finally, we address three tones for which identification accuracy was the most insensitive to contextual information. Among the multiple comparisons (Table 2.3), Tone 23 had the most nonsignificant results at the 0.05 level (6 of 6 comparisons), while Tone 21 had 3, as well as a nonsignificant comparison between the ISO and TRI conditions, and Tone 22 had 3.[5] We believe it is no accident that all three tones are in the crowded lower pitch range: Tone 23 and Tone 22 showed the lowest overall identification accuracy in connected speech (Fig. 2.3) and were confused with a mix of level and contour tones (Fig. 2.4). The significantly higher accuracy for both tones in isolation vs. in the TRI condition suggests that Tone 23 and Tone 22 are both simply confusable tones in connected speech, at least as the middle member of tritones with intial and final Tone 33s, so additional local context may be minimally informative for tonal identification.

Tone 21 stands out among the three tones because its overall accuracy rivaled that of Tone 55, the most accurately identified tone. The availability of voice quality cues beyond purely f0 value-based cues may have been a factor in the relatively context-independent perception of Tone 21, since Chapter 3 shows that the presence of creak can increase tonal identification accuracy and bias perception of Tone 21 for Cantonese listeners. Of all the tones we recorded, Tone 21

[5] Because the Tone 23 contours from speaker m5 were similar to his Tone 33 contours, we also checked the effect of CONTEXT when his stimuli were withheld. There was still only one significant multiple comparison, between MONO and TRI.

was most often produced with non-modal phonation, typically vocal fry (Gerratt and Kreiman, 2001). Tone 21 has previously been anecdotally noted to co-occur with creak (Vance, 1977; Matthews and Yip, 1994; Flynn, 2003). As a measure of the prevalence of creak in Tone 21 realizations, we counted the number of stimuli with at least 3 voiceless frames in RAPT f0 extraction. In isolation, 8 of the 14 such files were Tone 21 productions, about half of the Tone 21 stimuli, (3 were Tone 23 productions), and in the tritone stimuli, 4 of the 4 such files were Tone 21 productions.

2.4 Computational modeling

The perception experiment (§2.3) showed that the post-target syllable improves native listener tonal identification accuracy for Tone 25, while the pre-target syllable is minimally informative for Tone 25 identification, but is informative for the identification of level tones; it also showed that tones uttered in isolation can be identified as accurately as tones presented with preceding and following syllables from connected speech. While we suggested that these results could be explained with reference to the separability of the f0 contours of the tones conditioned on the different CONTEXT conditions, there are in fact an infinite number of acoustic parameters potentially available to the listener. Moreover, there is an unbounded range of influences outside the speech signal that the listener could bring to bear on the classification task, such as tonal change in reduplication, processing asymmetries for preceding vs. following context or for hearing two syllables in a row consisting of the same segment sequence (*wai wai* in the PRE condition vs. *wai mat* in the POST condition).

To support our claim that separability in an acoustic space defined by simple f0-based parameters could largely explain the listeners' behavior, we used computational methods to model the classification problem presented to the listeners, defined under precise assumptions. Our purpose was to determine: *given a minimal acoustic parameterization of the speech signal with only f0-based parameters and abstracting away from other sources of evidence, could the stimuli in our experiment be classified with results consistent with listener behavior: (i) with higher accuracy for Tone 25 identification with contextual parameters from the post-target syllable than from the pre-target syllable and similar confusion patterns, and (ii) with comparable overall accuracy for tones uttered in isolation and in full tritones from connected speech?* In the modeling, we defined the raw acoustic parametrization of the stimuli to be f0 values extracted from the stimuli with a 10ms frameshift, and because how finely listeners track the unfolding of the speech signal for tonal perception is not known, we also derived f0 parameter sets varying in temporal resolution that were sets of mean f0 values over 30-ms windows for 2, 3, 5, and 7 windows uniformly distributed over the syllable. For the 10ms frameshift, there were 23 f0 values sampled per syllable, 69 in total for modeling the tritone condition, 46 for each bitone condition, and 23 for the MONO condition. We tested and trained within the same condition, since the classifiers, being trained on a particular condition, are necessarily defined over the parameterization of the condition. We chose linear support vector machines (SVMs) as our classifiers (Vapnik, 1995; Cortes and Vapnik, 1995; Burges, 1998). SVMs are well-characterized mathematically, widely used in machine learning and have been used in automatic tonal recognizers, e.g. Levow (2005); Peng and Wang (2005).

Following Bennett and Bredensteiner (2000), we sketch a geometrical characterization of how SVMs work for a binary classification problem, e.g. for two tone classes, H or L. Each stimulus is parameterized as a real-valued p-dimensional vector and labeled as an H or L. Thus, the H and L stimulus sets each comprise a set of points in $\mathbb{R}^p$. The SVM algorithm determines an optimal decision rule to assign a class label to a stimulus. A linear SVM determines a $p-1$ dimensional separating hyperplane as a decision boundary in the parameter space, i.e. a 1-D separating line for stimuli parameterized in 2-D space, $\mathbb{R}^2$, whose direction is determined as a linear combination of the parameters: a parameter whose weight, called the primal weight, has a greater magnitude in the linear combination has a greater influence in deciding the classification of a stimulus.

The SVM algorithm defines the optimal separating hyperplane as the one maximizing the distance from the hyperplane to the H and L sets. This bisects and is orthogonal to the line segment between two closest points of the convex hulls of the H and L sets (Boyd and Vandenberghe, 2004, p. 46-49), where the convex hull of a set is defined as the set of points enclosed in the tightest rubber band one can stretch around the set. If the H and L sets are linearly inseparable, i.e. if their convex hulls overlap, then a soft margin SVM algorithm can be used. This allows for some points to be on the wrong side of the margin in determining the optimal separating hyperplane, and the tradeoff between maximizing the margin and minimizing classification error is balanced by tuning a soft margin parameter.

Since we desire the classification rule chosen by the SVM algorithm to generalize beyond the training data provided to the algorithm, evaluation of classifier performance proceeds by determining classification accuracy on test data, data not in the training data set.

2.4.1 Methods

The linear SVMs were implemented with LIBSVM (Chang and Lin, 2001). The SVM algorithm involves calculating Euclidean distances in the parameter space.

This means that it is necessary to scale the data, such that parameters with a greater range do not dominate the direction of the optimal separating hyperplane relative to parameters with a smaller range; it also necessary for the training and test data to be scaled in the same way. Thus, the parameter sets used were z-score standardized log-transformed f0 values rather than f0 in Hz (§2.2.3), cf. (Levow, 2006, §2.3).

LIBSVM treated the multiclass 6-way Cantonese tone classification problem in a standard way by decomposing it as $\binom{6}{2} = 15$ binary classification sub-problems and using a voting strategy to combine the 15 decisions. For each temporal resolution of f0 parameterization, (2, 3, 5, or 7 windows per syllable, or with 10 ms frames), for each CONTEXT condition for a given temporal resolution, 5-fold cross-validation was performed, and the data was partitioned into 5 folds, one fold per stimulus speaker. Rotating across the folds, a single fold (18 stimuli, 1 speaker) was withheld as test data, and the remaining four folds ($4 \times 18 = 72$ stimuli, 4 speakers) were used for training data. The soft margin parameter was chosen for each rotation using 5-fold cross-validation on the training data. All classification results discussed below are averaged across the results from the 5 rotations; standard error for classification accuracy is calculated from the variance of the accuracy over the 5 folds.

2.4.2 Results

Overall SVM classification accuracy aggregated over tones and speaker folds was not significantly different between the modeled POST and PRE bitone conditions for any sampling resolution, except for 2 windows/syllable, when accuracy in the POST condition was significantly higher ($t(4) = -6.32$ (paired by fold), $p = 3.20 \times 10^{-3}$). However, across context conditions within each temporal resolution, there

	2/syll	3/syll	5/syll	7/syll	10ms
MONO	33.00 (18.26)	53.33 (17.00)	66.67 (14.91)	66.67 (14.91)	66.67 (14.91)
PRE	40.00 (6.67)	46.67 (17.00)	66.67 (14.91)	73.33 (12.47)	66.67 (14.91)
POST	86.67 (13.33)	80.00 (13.33)	86.67 (13.33)	86.67 (8.16)	86.67 (8.16)
TRI	93.33 (6.67)	93.33 (6.67)	86.67 (8.16)	86.67 (8.16)	86.67 (8.16)

Table 2.6: SVM classification accuracy for Tone 25 in the connected speech conditions for each temporal resolution, from 2 to 7 windows/syllable, and also for a 10 ms frame shift. Accuracies and SE (in parentheses) were calculated over speaker folds.

were in fact almost no significant differences in classification accuracy at all—other than the one just mentioned, there were only significant differences between the tritone condition and the MONO and PRE conditions for 3 windows/syllable.

Like in the analysis of the perceptual experiment, we analyzed results by individual tones to unpack what insight we could gain from the computational modeling. Classification accuracy for the level tones were near-perfect independent of context or temporal resolution, especially for Tone 55 and Tone 33. For Tone 25, for every temporal resolution, classification accuracy in POST was higher than in PRE (Table 2.6). Misclassifications of Tone 25 were generally due to mislabelings as Tone 33 and Tone 22 in the PRE condition but Tone 23 in the POST condition. In addition, classification accuracy for Tone 25 in the PRE condition was close to that of the MONO condition, while accuracy in the POST condition was relatively higher, closer to that of the TRI condition for every sampling resolution. These trends are most apparent with the 2 windows/syllable sampling resolution, the only resolution setting in which t-tests, paired by fold, showed (near) significant differences where they were expected based on human results, between the bitone conditions, and between MONO and POST and PRE and TRI: t(4) = -0.34, p = 0.75 for MONO-PRE, t(4) = -1.73, p = 0.16 for MONO-POST, t(4) = -3.5, p = 0.025 for PRE-POST, t(4) = -6.5, p = 0.0028 for PRE-TRI, t(4) = -1, p = 0.37 for POST-TRI (p-values uncorrected for the 4 comparisons)).

Furthermore, the mean primal weights aggregated over folds for our models of the tritone condition showed the general trend that the weights defining separating hyperplanes for binary classification problems involving Tone 25 were of relatively higher magnitude for the post-target f0 values than the pre-target f0 values and target f0 values (Fig. 2.7). That is, for 2-way classifications for Tone 25 vs. each tone in the set of the other five tones, except Tone 55, the classification decision was more strongly influenced by post-target acoustic information than pre-target and target syllable acoustic information.

SVM classification accuracy for the isolated stimuli was strikingly low for Tone 21, with accuracy ranging from 6.67-46.67% over the different temporal resolutions, quite unlike the high Tone 21 accuracy in human perception. Despite this, there was no significant difference between overall classification accuracy for the isolated stimuli in comparison with the tritones from connected speech, in t-tests paired by fold for each temporal resolution; for 7 windows/syllable, results were closest to being significant, with accuracy higher for the tritones, $t(4) = 2.49, p = 0.061$. Tone 23 classification accuracy was not strikingly higher for the isolated stimuli than for other CONTEXT conditions as in human perception; rather, it generally hovered between 40-60% regardless of condition, and in the isolation condition, Tone 23 was frequently misclassified as Tone 21 and less often as Tone 22, while in the other conditions, Tone 23 was misclassified as Tone 22 more often than as Tone 21.

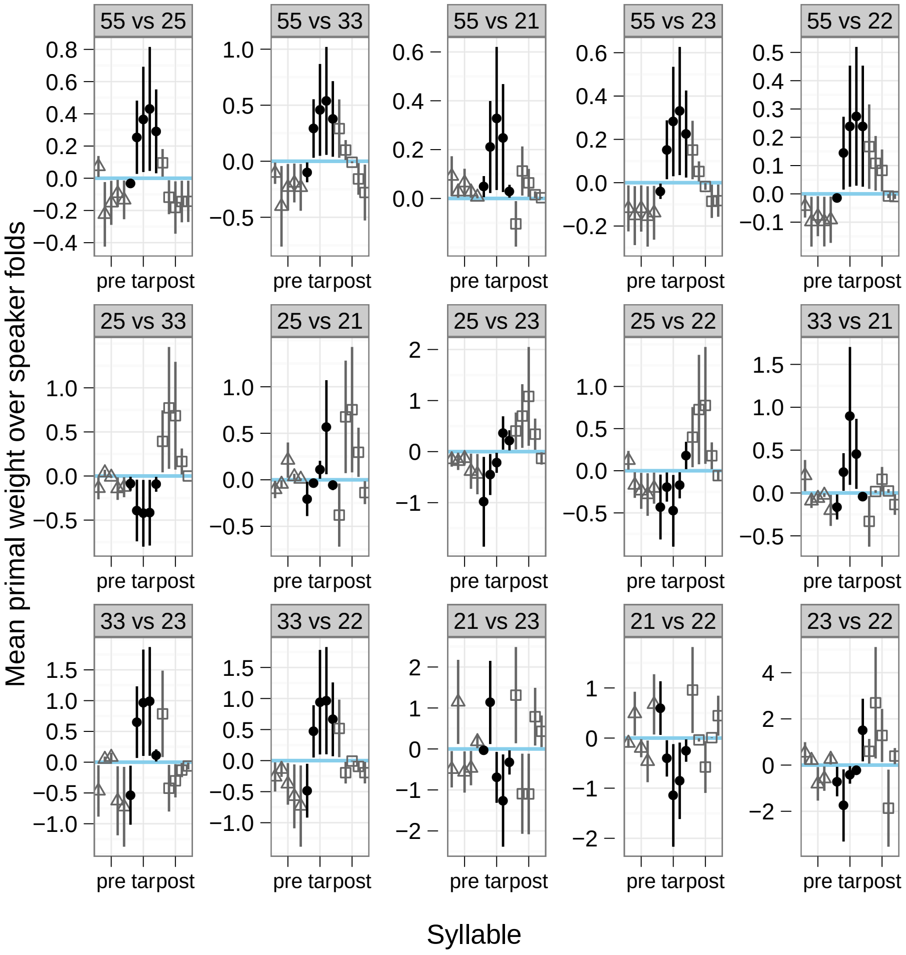


Figure 2.7: Mean weights aggregated over folds for defining separating hyperplanes for each binary classification problem decomposed from the 6-way Cantonese tone classification problem, for the tritone condition and a temporal resolution of 5 windows/syllable for parameter extraction. The weights (±1SE) are indicated for the f0 parameters from the pre-target syllable (grey triangles), the target syllable (black circles), and the post-target syllable (grey squares), sorted by temporal order.

2.4.3 Discussion

Given only f0-based parameters and abstracting away from other sources of evidence, the stimuli in our experiment showed a trend for higher classification accuracy with support vector machines for the Tone 25 rise with contextual parameters from the post-target syllable than from the pre-target syllable. Confusion patterns were similar to those in the human perception experiment, where Tone 25 was more confusable with level tones when the pre-target contextual information was available. Furthermore, the relative magnitudes of weights of the parameters defining separating hyperplanes for binary classification decisions involving Tone 25 in our models of the tritone condition were generally higher for the post-target parameters than the target and pre-target parameters. Classification accuracy for stimuli uttered in isolation was comparable with accuracy for tritone stimuli from connected speech. All these results were robust across the different temporal resolutions for parameter extraction, from 2 windows/syllable to 7 windows/syllable, and with 10 ms frame shifts. Thus, overall, the computational modeling using f0-based parameters supported some of the main claims of our acoustic account of listener behavior.

However, the computational modeling was also limited in substantial ways. First, the tiny data set—designed to be suitable for a human experiment— made the machine results less sensitive: we were only able to test the trained classifiers on 3 examples per tone, and thus, few results comparing contexts reached statistical significance. This is why we did not test for a negative effect of CONTEXT on Tone 23 classification accuracy.

Second, the scaling of the data in pre-processing, while standard in computer science and statistics, implies an assumption of a cognitive model where the listeners scaled the data in such a way that it was perceived as normalized with respect to each speaker. This is a methodological abstraction, but nevertheless an assumption that seemed to greatly change the setup for the classification problem relative to what the listeners were doing. None of the human perception results for level tones were reflected in the machine results: level tones were classified with near-perfect accuracy regardless of any conditioning, while contextual information, especially pre-target context, greatly improved level tone accuracy for humans. The most plausible explanation for this is that the z-score normalization for the machine classification muted the informativity that syllable-extrinsic context provided for perceptual normalization. In support of this explanation, the weights for the separating hyperplanes in Fig. 2.7 show a very limited influence of the pre-target contextual information overall, and for level tones, the highest weight magnitudes are assigned to the target syllable parameters. The computational modeling for the rises seemed to be much more human-like, and this was probably because contour tone perception for human and machine relied less on normalization.

Third, the parameterization of the stimuli poorly captured the voice quality information available to the listeners, cf. §2.3.3. While human identification of the Tone 21 fall was around 70-80% accurate overall, with little effect of context, machine identification was typically 30% or below for isolated stimuli, but around 60-70% for the other contexts. We conjecture that the misfit between human and machine is due to the lack of voice quality information beyond smoothed f0-based values in the computational modeling : in particular, because it is difficult to deal with missing values in machine learning, we filled the frames where RAPT assigned no f0 values (typically in creaky regions) using linear interpolation. For the isolated stimuli, in particular, where nearly half the Tone 21 stimuli had creaky regions and missing frames, our crude smoothing clearly did not capture how listeners perceive creak, and also resulted in confusion patterns quite different from human confusion patterns, such as Tone 23 being highly confusable with Tone 21 in isolation for machines, while Tone 23 was highly confusable with Tone 25 in isolation for humans. While one could argue that with more sophisticated heuristics, we could have presented corrected f0 tracks to the algorithm that may have increased Tone 21 accuracy, the fact remains that modeling human tonal perception using purely (smoothed) f0-based parameters does not capture what humans are doing.

2.5 General discussion

In this study, we found evidence from a Cantonese tonal perception experiment that listeners use following context in tonal identification: in speaker-independent tonal perception of tones extracted from connected speech, listeners were more accurate in identifying the Tone 25 rise and Tone 21 fall when they were also provided the syllable following the rise, than when they were also provided the syllable preceding the rise. The preceding and following syllables improved listener accuracy on complementary sets of tones; in general, while the post-target syllable improved contour tone accuracy, the pre-target syllable improved level tone accuracy. Computational modeling of the stimuli using support vector machine classification of standardized log-transformed f0 parameters supported the idea that following context benefited Tone 25 perception because of peak delay. With the purely acoustic information available, machine classification showed a trend for higher accuracy with the post-target parameters available than with the pre-target syllable parameters available, and classification decision rules involving Tone 25 generally weighted post-target parameters more highly than pre-target or target syllables parameters.

Furthermore, tones were more perceptually distinct for the listeners in isolation than in connected speech, but only when there was no local acoustic context available, and even in this case—the MONO condition—Tone 55 identification accuracy was higher for connected speech probably due to carryover coarticulation.

When the tones were presented with the pre- and post-target syllables as tritones, overall tonal identification accuracy was comparable with that in isolation. As we expected, the perception of level tones suffered in isolation relative to when context was available in connected speech. Surprisingly, the low level tones and rises, Tone 23 and Tone 22, were perceived with higher accuracy in isolation than in the tritone condition. This was most likely because of the distinctiveness of rises and levels in isolation: rises were confused with rises, and levels with levels, while this was not the case in connected speech, especially in the crowded lower pitch range, either for the listeners or in computational modeling.

Therefore, tones in isolation in our experiment were more perceptually distinct than in connected speech, even in speaker-independent perception, in a way very relevant for understanding phonological representation: confusion patterns in isolation patterned along distinctions drawn by standard phonological features, e.g. Wang (1967), which divide levels and rises and falls. However, tones in isolation were not identified by listeners more accurately than tones in connected speech—provided that listeners were provided with local acoustic context, the neighboring syllables. This result, coupled with our finding that both preceding and following context improve tonal recognition for listeners and the results from the literature that show improvement in automatic tonal recognition with preceding and following context as well, suggests that tonal representations can span a domain beyond the associated syllables. Alternations between citation tones and tones in connected speech may involve mapping to tones in connected speech *including contextual features from neighboring syllables.* Furthermore, we suggest that isolation can be treated as one of many contexts rather than only as a privileged base. With these suggestions, we mean to address the cognitive representation of tones, and not dismiss the utility of parameterizations for tone that may very well achieve high rates of recognition in automatic tonal recognition.

Based on our perception study and the automatic tonal recognition studies using contextual parameters referenced in §2.1, it seems that only a limited amount of contextual information may need to be included in tonal representation, e.g. just a few samples of f0-based parameters from each neighboring syllable. Adding contextual information, when it is as limited as we are proposing, does not explode the complexity/dimensionality of the representation. Moreover, contextual information seems to be necessary to account for listeners being undaunted in the face of extreme allophonic variation (from the perspective of citation underlying forms) *so long as the neighboring syllables conditioning the variation are provided* (see examples in Shih and Kochanski (2000)). Xu (1994) found that in such cases, such as when Mandarin Tone 4, a fall in citation form, appears as a rise when flanked by a preceding tone ending low and a following tone starting high, tonal identification accuracy dropped below chance when the preceding and following syllables were replaced with white noise, while it was 97% with the original flanking syllables.

From a typological and learnability perspective, both the ideas of tonal representations spanning a temporal domain beyond a single syllable and isolation as one among many contexts are natural ones. First, tones seem to be different from other phonological categories like vowels or consonants because it is the norm for them, much more than vowels or consonants, to spread and shift beyond the syllables they are associated to (Goldsmith, 1976). Peak delay, for instance, is a typical process, as is tonal shift. Tonal representations spanning beyond the associated syllable of a tone can help explain diachronic processes in which phonetic peak delay could result in tonal shift (Kaplan, 2008).

Second, cross-linguistic field work has shown that in some languages, neutralizations of tonal contrasts occur when tones are next to other tones in connected speech in sandhi contexts; in others, neutralizations occur in isolation: "whether one must take the citation tone or the sandhi tone to be underlying depends on the observed patterns of tonal alternations" (Chen, 2000, 51). This is not an observation special to tones; it is a general and classic one in phonology, as discussed in Kenstowicz and Kisseberth (1977, 18) and Hayes (2008, 165). The language specificity of which context might provide an underlying form implies that this is something that must be learned. From a learnability standpoint, too, the task of learning to map tones in connected speech devoid of context with isolated forms gives rise to a challenging learning puzzle that is manageable if we consider mapping to forms including (a small amount of) context.

Further explorations of tonal representations spanning multiple syllables using computational modeling would benefit from a better understanding of how to model the scaling of parameters in a human-like way, as the z-score standardization used here was the likely culprit for poor fit with listener behavior in integrating information from neighboring syllables for level tone identification. In addition, it would be interesting to find a cognitively-motivated way to model training based on parameters from one particular context, but testing on parameters from another. This would help us understand mapping between alternations. However, it is non-trivial to propose cognitively-motivated rules for how to adapt a classifier defined in one parameter space to classify objects in another parameter space.

Further explorations using perceptual experimentation could consider alternatives to the stimuli design here, where contexts from connected speech were modeled using extracted sentence-medial tritones. Systematic errors consistent with expectations of stimuli-final downtrends suggests that it may be difficult for listeners to treat such extracted stimuli as fragments of connected speech.

While surely more cross-linguistic work in typologically different tonal systems would be very informative, Cantonese presents an interesting problem for alternation between allophonic variants of tones. Mok and Wong (2010a,b) have studied tonal mergers in Cantonese, between Tone 25/Tone 23, Tone 33/Tone 22, and Tone 21/Tone 22 in present-day Cantonese. In post-hoc analysis of our listeners, we checked for these perceptual mergers in our listeners in the isolation condition by checking if a majority of responses to a member of those merger pairs was the other member of the pair. Among the merger pairs Tone 33/Tone 22 and Tone 21/Tone 22, only one subject confused Tone 22 with Tone 21 more than 53% of the time; otherwise, no subject showed confusability among the merger pairs above 50%. However, there were 9 subjects who confused Tone 25 with Tone 23 at least 50% of the time, one 87% of the time, and the rest 73% of the time or less. Did these subjects have Tone 25/Tone 23 mergers? Not if mergers are unconditioned on alternation between allophonic variants: in the TRI condition, 4 subjects had above 50% confusability (53-60%) of Tone 25 with Tone 23, but none of the subjects were the same as the 9 subjects who performed poorly in isolation. For our experiment, we abstracted away from tonal merger in present-day Cantonese, but a question that one could explore with subjects with tonal mergers is how the tonal mergers interact with alternation and allophonic variation.

CHAPTER 3

The role of creaky voice in Cantonese tonal perception[1]

Glottalization: With loud stress, /^{3}/ and /^{5}/ often have glottal friction during the lowest-pitched phase of the contour.

Charles Hockett on Peiping phonology, (Hockett, 1947, 256)

Other elements, such as a slightly shorter duration of the 2nd Tone and slight vocal constriction at the trough of the 3rd Tone, may be considered as secondary, though they may become important under special conditons, such as in whispered speech.

Yuen Ren Chao on Mandarin tones, (Chao, 1956, 53)

3.1 Introduction

There are two named typological patterns for languages in which variation in the voice source over the vowel/rime results in contrasts in lexical meaning. Tone languages are traditionally defined to be languages where pitch is lexically contrastive (Yip, 2002). Register languages, at least in opposition to tone languages, are traditionally defined to be languages where phonation is lexically contrastive

[1] **This chapter represents joint work with Hiu Wai Lam. Experiment 1 was previously described in her bachelor's honors thesis.**

(Henderson, 1952): “phonation type is to a register language what tones are to a tone language” (DiCanio, 2009, p. 162).[2]

In addition, some languages have both contrastive pitch and contrastive phonation. In Jalapa Mazatec, breathy, modal, and creaky phonation can occur on high, mid, and low tones for a three way-by-three way set of suprasegmental contrasts (Kirk et al., 1993; Silverman et al., 1995; Garellek and Keating, 2011). Other such languages include Tibeto-Burman languages such as Mpi (Ladefoged and Maddieson, 1996; Blankenship, 2002), Jingpho (Maddieson and Ladefoged, 1985), and Yi family languages (Kuang, 2011), some of which may have phonation contrasts for only a subset of the tones.

Hmong and Vietnamese are traditionally called tone languages, but they also are of this type. In Hmong, one of the falling tones is consistently breathy (Huffman, 1985; Andruski and Ratliff, 2000), and in dialects of Vietnamese, multiple tones are consistently realized with different kinds of nonmodal phonation (Pham, 2003; Michaud, 2004). For these languages, acoustic studies suggest that phonation is contrastive (Andruski and Ratliff, 2000; Pham, 2003), i.e. fundamental frequency (f0) values alone are insufficient for defining a contrast, and perceptual studies confirm that listeners use voice quality cues in tonal perception (Andruski, 2006; Brunelle, 2009; Kirby, 2010).

There are also languages in which, between pitch and phonation, only one is contrastive, but the other, though non-contrastive, is conditioned on the contrastive dimension. The definition of register languages, in fact, traditionally

[2] The literature on voice quality is fraught with inconsistencies in terminology (Gerratt and Kreiman, 2001; Surana and Slifka, 2006). We use *phonation* as a term for a specific class of *voice quality* which can be divided into two classes: *modal* (default, baseline) and *nonmodal phonation.* The class of nonmodal phonations discussed in this paper is sometimes called *creaky* in referring to the acoustic signal, and we use *creaky/creak* to describe both period doubling and vocal fry. We reserve the term *creaky voice* to refer to the percept associated with *creak* in the acoustic signal.

encompasses more than just contrastive phonation: while pitch is not considered to be the "primary relevant feature", tendencies for pitch characteristics are standardly used in defining a register in a register language (Henderson, 1952, p. 151). Descriptions and phonetic studies of register languages do not abstract away from studying how f0 and pitch fit in the definition of registers (e.g. Abramson et al. (2004) on Suai and DiCanio (2009) on Chong); in fact, though register languages are traditionally defined to have contrastive phonation, results from perception experiments in register languages indicate that it is not clear that either pitch or phonation alone is contrastive (Abramson et al., 2004; Gruber, 2011).

In contrast to in register languages, in tone languages, it is the norm to abstract away from noncontrastive phonation cues that may be conditioned on tonal categories. Yet, in some tone languages, it is well known that one or more tone categories may inconsistently be realized with nonmodal phonation. The standard example for this pattern is Mandarin: Tone 3 (214, ˨˩˦), the lowest tone in the inventory, is sometimes creaky (Hockett, 1947; Chao, 1956; Gårding et al., 1986; Klatt and Klatt, 1990; Davison, 1991; Belotel-Grenié and Grenié, 1997). Because the presence of creak in Tone 3 productions within and across speakers is variable, and implicitly since perception experiments as well as automatic tonal recognition studies typically abstract away from creak for methodological reasons (e.g. Whalen and Xu (1992, p. 27-29); Zhang and Hirose (2004); Wang et al. (2010)) f0-based features are assumed to be sufficient for discrimination of Tone 3 from the other tones of Mandarin.

Another language of this type is Cantonese, which has six tones: high level (Tone 1, 55, ˥), high rising (Tone 2, 35/25, ˧˥/˨˥), mid level (Tone 3, 33, ˧), low falling (Tone 4, 21/11, ˨˩/˩), low rising (Tone 5, 23/13, ˨˧/˩˧), and low level (Tone

6, 22, ˨) (Matthews and Yip, 1994).[3] Cantonese has anecdotally been reported to have an (inconsistently) creaky Tone 4 (Vance, 1977, 105; Matthews and Yip, 1994, 22), and Yu (2010) confirmed this in a small corpus of Cantonese data. Santa Ana Del Valle Zapotec (Oto-Manguean, Mexico) has also has been found to have creak inconsistently realized on a falling tone (Esposito, 2003, p. 39), dependent on prosodic position.

This paper addresses the role of phonation cues in tonal perception in Cantonese, an exemplar of a tone language with non-contrastive phonation cues. An initial question to ask of such languages is: *how prevalent is the presence of creak and is it conditioned by tonal category?* Small corpus studies of Mandarin and Cantonese connected speech suggest that creak is prevalent in the lowest tone of the inventory and that its presence is conditioned on tonal category because creak is less frequent in other tones, cf. Table 3.1; creak is not uniformly distributed among the tones.

Since the presence of creaky voice in Mandarin and Cantonese is both frequent and conditioned on tonal category, a natural research question is: *are listeners sensitive to creaky voice in tonal perception in tonal languages with non-contrastive phonation cues?* While a number of perceptual studies have shown that listeners of tonal languages with contrastive phonation use phonation-based cues in tonal perception, there are few such studies on languages with non-contrastive phonation cues. Gårding et al. (1986) used a 2-way forced choice tonal identification task for Tone 3 and 4 in Mandarin and tested listeners on a continuum of timepoints for a minimum (turning point) in the f0 contour. They compared the identification curve for stimuli without creak, and those resynthesized to be creaky by introducing pitch halving in the middle of the vowel. They

[3] Some descriptions also distinguish these tones from the shorter entering tones (high, mid, and low level) which occur in syllables with unreleased stop codas.

		Frequency of creak (%)		
Study	Data	Overall	Lowest tone	Other tone
Belotel-Grenie and Grenie [1994, 1997]	Mandarin, lab speech, 7 speakers (4M/3F)	32.3	78.4	32.1 (T4, 51)
Belotel-Grenié and Grenié (2004)	Mandarin, newscast speech, 1 speaker (1F)	9.4	26.5	3.3 (T4, 51)
Yu (2010)	Mandarin, lab speech, 8 speakers (4M/4F)	20.8	68	8 (T4, 51)
Yu (2010)	Cantonese, lab speech, 8 speakers (4M/4F)	4.5	25	2 (T2, 35/25)

Table 3.1: Small corpus studies of presence of creak in Mandarin and Cantonese. The frequency of creak is given: over all tones, for the lowest tone in the tonal inventory (Tone 3 in Mandarin, Tone 4 in Cantonese), and for the non-lowest most frequently creaky tone. The discrepancy between the prevalence of creak in Mandarin between laboratory speech and newscast speech is striking, but may be due to individual and/or speech genre variation in use of creak.

found that the pitch halving had very little or no effect on tonal identification. However, Belotel-Grenié and Grenié (1997) performed a tonal identification gating task in Mandarin with both creaky and non-creaky Tone 3 natural stimuli and found that the recognition point for the listeners came sooner for creaky Tone 3s. Donohue (2011); Yang (2011) also reported evidence for the sensitivity of listeners to creaky voice in Fuzhou and Mandarin tonal perception, respectively. No previous work has demonstrated that the presence of creak can improve tonal identification accuracy in a language with non-contrastive phonation.

In this paper, we performed two experiments in Cantonese tonal perception to build on these previous studies. The first experiment was an initial test for sensitivity of listeners to creaky voice in Cantonese tonal perception. It was a 6-alternative forced choice tonal identification task of Cantonese monosyllables extracted from a corpus of multispeaker connected speech. Drawing on the natural variation in the corpus, we chose half of the Tone 4 stimuli to be creaky and the other half non-creaky. Beyond hypothesizing a sensitivity to creaky voice in Cantonese tonal perception, we hypothesized that identification accuracy for the creaky Tone 4s would be higher than that of the non-creaky Tone 4s, and reaction time for the listener response would be shorter for the creaky Tone 4s than non-creaky Tone 4s.

Experiment 1, like previous experiments on the role of creaky voice in tonal perception such as Gårding et al. (1986); Belotel-Grenié and Grenié (1997), treated creaky voice as a single cue with no internal structure. This is because our current understanding of how creak is generated by the voice source and how listeners perceive variability in the realization of creaky speech is still very limited. Thus, the second experiment elaborated on our knowledge of *how* listeners use creak in tonal perception. It was designed to follow up on Experiment 1 and bear on the following questions: (i) *How do listeners integrate creaky voice with pitch cues?* and (ii) *Are listeners sensitive to details of creak?* The experiment was a controlled 2-alternative forced choice tonal identification task between the low fall Tone 4 and the low level Tone 6, the tone most confusable with Tone 4 (Ma et al., 2005; Khouw and Ciocca, 2007; Fok, 1974). Level f0 Tone 6 productions from a male and female speaker were resynthesized and cross-spliced with different natural creaky productions of Tone 4 and presented to the listeners. Additionally, the tone to be identified was preceded by a syllable whose f0 was shifted up and down to see if listeners integrate creak with contexual f0 information.

The results of both experiments bear on a range of issues for tonal languages with non-contrastive phonation cues: (1) how tones are represented in the acoustic space relevant for human tonal perception, i.e. if there is reason to define tones in an elaborated space with voice quality parameters beyond f0 values, and (2) if automatic tonal recognition might benefit from changing current approaches which abstract away from creak in the speech signal.

In the rest of this paper, we report on Experiment 1 in Sec. 3.2 and Experiment 2 in Sec. 3.3 and conclude with a general discussion (Sec. 3.4).

3.2 Experiment 1: Cantonese tonal identification and creaky voice in Tone 4

Previous results regarding whether listeners are sensitive to creaky voice in tonal perception in tonal languages with non-contrastive phonation cues are limited in number and inconsistent, cf. Sec. 3.1, and primarily only Mandarin has been studied. Experiment 1 tested whether Cantonese listeners showed sensitivity to creaky voice in tonal perception.

We designed the tonal identification task to be very difficult to maximize the chance that listeners could benefit from creaky voice in tonal perception, since the previous studies in Mandarin have had inconsistent results, and since the listeners in Belotel-Grenié and Grenié (1997) were at ceiling for tonal identification accuracy so that it was not possible to detect the effect of creaky voice on accuracy. The realization of tones in the stimuli was highly variable since the stimuli were extracted from a multispeaker corpus of connected speech designed to exemplify contextual tonal variation. The contextual information actually available to the listener was minimal because the stimuli were monosyllables, the presentation of stimuli was randomized, and the identification task required choosing between all six tones of Cantonese.

3.2.1 Methods

3.2.1.1 Materials

The stimuli were 596 tokens of sentence-medial /lau/ syllables drawn from a Cantonese tonal production corpus consisting of sentences [lei$^{25/35}$ jiu^{33} lau lau jak$^{\neg 33}$ tʃoeŋ^{33}/kap$^{\neg 33}$/sou^{33}] ‘*you want Lau-Lau to eat the sauce/pigeon/vegetable*’ with the target bitone /lau-lau/ over all 36 possible combinations of tones T1 to T6.[4] For all tones, *lau* is a real word, although *lau*^{33} is uncommon. There were 108 sentences in total, 5 repetitions of each sentence, and 15 speakers in this corpus. From this corpus, 72 examples were drawn in total for each tone from 4 male and 4 females who were chosen to be widely distributed in pitch range.

[4] Reduplicated expressions, e.g. when the bitone is composed of a sequence of identical morphemes, can be the target of tonal change in certain morphological/semantic contexts, such as for marking vocatives or intensification in adjectives (Matthews and Yip, 1994; Yu, 2007, 2009), but there were no tonal changes in the recordings.

The bitone sequence and position of the syllable in the bitone (1st or 2nd) were balanced for each speaker. For the T4 tokens, half of the tokens were chosen to be creaky and the other half non-creaky. By non-creaky, we do not mean modal. Non-creaky tokens may have had a region of relatively low amplitude or breathiness. Because of great interspeaker variability in the prevalence of creaky T4s, it was not possible to fully balance the presence of creak in T4s. All female speakers contributed 6 creaky and 6 non-creaky T4s, but one male speaker contributed only 2 creaky tokens and another 9.

T4 tokens were determined to be creaky by listening and manual inspection of the waveform and spectrogram in Praat (Boersma and Weenink, 2010). A token was defined to be creaky if it had the auditory percept of creaky voice, as determined by the authors and if: 1) there were alternating cycles of amplitude and/or frequency or irregular glottal pulses in the waveform, 2) missing values or discontinuities in the f0 track determined by Praat's autocorrelation algorithm with default settings, and/or 3) the appearance of strong subharmonics or lack of harmonic structure in the narrowband spectrogram. Generally these three indicators occurred simultaneously. An expert in voice quality listened to the creaky/non-creaky subsets and confirmed that the tokens had/didn't have the percept of creaky voice. All tokens were resynthesized using PSOLA in Praat to have equal average amplitude, and the duration of each token was equalized to 313ms, the grand mean of token durations. Duration and amplitude were controlled since this study was designed to single out the contribution of creaky voice as a cue.

3.2.1.2 Participants

The participants were 16 native Cantonese speakers recruited from the student population at the University of California, Los Angeles; all spoke Cantonese on a daily basis. They received cash compensation. All were born in Hong Kong except one born in Macau, and their mean time of stay in the US was 4.0±1.8 years. There were 11 males (age 20.6±1.6 years) and 5 females (age 21.2±0.8 years). Two other participants were tested but omitted from analysis due to equipment failure.

3.2.1.3 Procedure

Participants were tested in a sound-treated booth. The perception experiment was run in Matlab using Psychophysics Toolbox extensions (Pelli, 1997; Brainard, 1997). Stimuli were played from an Echo Indigo IO sound card on a laptop over studio monitor headphones at a standardized, comfortable volume, and the responses and reaction times of the subjects measured from the onset of the stimulus were recorded. The interstimulus interval was 3s.

The task of the participants was to identify each stimulus by a keyboard press of one of six keys labeled with the characters for the minimal tonal set over *lau*. The lexical meanings of the orthographic characters for Tones 1-6, respectively, were: ‘angry’, ‘twist’, ‘instigate’, ‘stay’, ‘willow’, and ‘leak’. Participants were asked to respond as quickly and accurately as possible and told they would be timed. They were also told that the stimuli were extracted from sentences *lei*$^{25/35}$ *jiu*^{33} *lau lau yak*^{3} *sou*^{33} and that the sentences were read by multiple speakers. The order of the stimulus presentation as well as which key was labeled with which word was randomized across participants, and participants received 3 short breaks during the experiment, which took about 45 minutes.

3.2.1.4 Data analysis

Statistical analysis was performed in R (R Development Core Team, 2010), using the ggplot2 package for creating graphics (Wickham, 2009). The overall confusion matrix for all stimuli was calculated and further analysis was performed on the T4 stimulus subset. Data were excluded from analysis for one T4 stimulus which sounded highly unnatural after resynthesis and yielded long reaction times that were outliers. The T4 stimulus subset was analyzed using mixed effects regression implemented by the lme4 package of Bates and Maechler (2010), a statistical method that is now prevalent in language research (Baayen, 2008). Mixed effects models, unlike ANOVAs, allow a unified treatment of continuous and categorical variables, are robust against unbalanced/missing data (present in this experiment), are robust against violations of homoscedasticity and sphericity, and allow the inclusion of complex random effects structure (Baayen et al., 2008; Baayen, 2008; Quené and van den Bergh, 2004). Logarithmically-transformed reaction times were analyzed using linear mixed effects regression.[5] Response accuracy for T4 stimuli was analyzed with mixed effects logistic regression since the dependent variable was binary: correct or incorrect. Linear models were fit using restricted maximum likelihood and logistic models with Laplace-approximated maximum likelihood (Pinheiro and Bates, 2000).

Forward model selection was used to test the partial effect of CREAK (present, absent) on correctness of T4 response and log reaction time while controlling for other variables: SYLLABLE (S1, S2), SPEAKER SEX (male, female), and SPEAKER (8 levels). SYLLABLE indicates whether a stimulus was uttered as the first (S1) or second (S2) syllable in the bitone; SPEAKER SEX refers to the sex of the

[5] Log-transforming reaction times is standardly performed to help satisfy the assumption of a normally distributed dependent variable in linear regression, although researchers have also argued for other transformations, e.g. Kliegl et al. (2010).

speaker of a given stimulus. SPEAKER was included as a non-interacted fixed effect to control for variability due to the different speakers who uttered the stimuli. Each of the fixed effects was a categorical variable and coded using numeric indicator variables (1,0); they were then mean centered to reduce collinearity and for model interpretability, e.g. for balanced binary variables, contrasts were set as -0.5 and 0.5 (Gelman and Hill, 2007, Ch. 4). Random intercepts for SUBJECT (the listener) were included, providing individual by-listener adjustments for T4 response correctness or log reaction time.

Successive nested models in the forward model selection were compared using likelihood ratio tests. Model likelihood is the probability of the data given the estimated model parameters, and for large datasets, differences in deviance ($-2\log(likelihood)$) between nested models fit to the same data by maximum likelihood approximately follow a χ^2 distribution.[6] Thus, χ^2 tests, with degrees of freedom corresponding to the difference in the number of parameters between two models under comparison, were used to test for significant improvement in fit to the data while penalizing for model complexity ($p < 0.05$) (Baayen, 2008, p. 253).

3.2.2 Results

Overall, identification accuracy for T4 was high (70.51%, SE=10.93%) compared to that of other tones, and the tone most confusable with T4 was T6, cf. Table 3.2.[7] Breaking down the 70.51% identification accuracy for T4 by phonation,

[6] For linear mixed effects models, models were refit for model comparison using maximum likelihood rather than restricted maximum likelihood.

[7] The identification accuracy of the other tones is not the focus of this paper, but we note the following: first, perhaps because T3 (*lau*^{33}) is uncommon, its identification accuracy was the poorest (32.23%). Perhaps because of that and also because most of the Cantonese level tones are towards the bottom of the pitch range, T1 was identified most accurately (85.94%). Also, there was a bias for T5 responses: 55.99% of T2 stimuli were identified as T5s, replicating

	Tonal identification response (%)					
Actual	T1	T2	T3	T4	T5	T6
T1	**85.94**	0.85	8.53	0.07	1.04	3.58
T2	1.30	**35.09**	3.65	1.50	55.99	2.47
T3	27.28	1.04	**32.23**	3.39	3.12	32.94
T4	1.89	3.19	2.93	**70.51**	8.07	13.41
T5	2.34	9.77	7.68	5.34	**63.93**	10.94
T6	8.46	1.76	20.25	15.56	4.56	**49.41**

Table 3.2: Overall confusion matrix for tones in Exp. 1

identification accuracy for T4 was 82.03% (SE=2.27%) for creaky T4s but only 58.98% (SE=3.57%) for noncreaky T4s. This boost in T4 identification accuracy due to the presence of creak was significant, as described below.

Model comparison (Table 3.3) showed that the best model for the probability of correctly identifying T4 with random intercepts for listeners included the following fixed effects: CREAK, SYLLABLE, SPEAKER SEX, and the interaction SYLLABLE:SPEAKER SEX.

All coefficients in the model were significant at the 0.05 level, cf. Table 3.4. Overall, the presence of creak significantly increased the probability of correct identification of T4, and the probability of correct identification was significantly higher for syllable 1 stimuli than syllable 2 stimuli from female speakers.

There was a nonsignificant trend of shorter reaction times for correct T4 identification of syllable 2 stimuli when creak was present. We performed model comparison of mixed effects linear models, with LOG REACTION TIME as the dependent variable and the same fixed and random effects as for the logistic models

the pattern of results for sentence-medial tones in Ma et al. (2005).

Fixed effects included in model				
SYLL	SEX	CREAK	SYLL:SEX	LL ratio test
✓	✓			
✓	✓	✓		$\chi^2(1) = 17.02, p = 2.0 \times 10^{-4}$
✓	✓	✓	✓	$\chi^2(1) = 13.20, p = 2.8 \times 10^{-4}$

Table 3.3: Summary of comparison of mixed effects logistic models for T4 identification correctness, including the fixed effects that were compared and the results of log-likelihood ratio tests for each successive comparison.

	Coefficient	SE	Z	p
(Intercept)	1.10	0.16	6.9	<.0001
CREAK	1.29	0.13	10.0	<.0001
SYLL	−0.50	0.13	−4.0	<.0001
SPEAKER SEX	−0.30	0.13	−2.2	<.05
SYLL:SPEAKER SEX	0.91	0.25	3.6	<.001
SPEAKER	0.14	0.03	5.1	<.0001

Table 3.4: Summary of fixed effects for mixed effects logistic model of correctness of T4 identification

discussed above. A log likelihood ratio test supported adding CREAK:SYLLABLE to a model with the fixed effects CREAK, SYLLABLE and SPEAKER, $\chi^2(1) = 3.50, p = 0.062$. While this was not significant at the 0.05 level, it was due to faster reaction time for creaky syllable 2 stimuli (mean log reaction time 0.41/[s] (SE 0.05)) than noncreaky ones (mean log reaction time 0.45/[s] (SE 0.04)).

3.2.3 Discussion

Experiment 1 demonstrated that the presence of creak significantly improved identification accuracy of Cantonese Tone 4, confirming our hypothesis about identification accuracy. There was also a trend for shorter reaction times in the presence of creak for a subset of the stimuli, the stimuli drawn from the second syllable of the target bitone in the recorded corpus, providing support for but not confirming our hypothesis for reaction time.

To our knowledge, this is the first experimental result suggesting that native speakers of a tone language with phonation as a non-contrastive cue for tone show improved tonal identification accuracy due to phonation cues. Belotel-Grenié and Grenié (1997)'s results showed an earlier isolation point for Mandarin T3 when creak was present, but could not show that listeners improved tonal identification accuracy for T3 because the listeners' performance was at ceiling in their task. Because our task was designed to be difficult—with randomly ordered multispeaker stimuli drawn from varying tonal contexts in connected speech and a larger, more confusable tonal inventory (six response choices rather than four, including three level tones and two rises)—the performance of listeners was not at ceiling.

There was also a significant effect of the interaction between SYLLABLE and SPEAKER SEX on tonal identification accuracy: syllable 1 stimuli were identified with higher accuracy than syllable 2 stimuli for the stimuli from female speakers. Post-hoc acoustic/auditory inspection of individual files suggested that for the female speakers, the syllable 1 stimuli included many strikingly canonical exemplars of T4, i.e. exemplars not heavily co-articulated with neighboring tones. This may have been partly because carryover (left-to-right) tonal co-articulation has been shown to be stronger than anticipatory (right-to-left) in Cantonese (Flynn, 2003; Wong, 2006), so that the variability in tones following syllable 1 in the recorded productions affected the realization of syllable 1 less than the variability in tones preceding syllable 2 affected the realization of syllable 2. This may have been why tonal identification accuracy was high for syllable 1 stimuli from females, especially for noncreaky syllable 1 stimuli (M=79.17%, *SE*=4.17) compared to noncreaky syllable 2 stimuli (M=53.65%, *SE*=4.16). Such variability in the stimuli also probably contributed to the restriction of the trend for shorter reaction times for creaky T4s to the second syllable of the recorded bitone. Listeners may have been closer to being at ceiling for stimuli drawn from the first syllable for noncreaky stimuli, so that the additional cue of creaky voice had a smaller effect on reaction times for the first syllable than the second.

The results from Experiment 1 suggest that listeners in Cantonese are aided in identifying T4 if creak is present. However, the stimuli used were drawn from naturally produced speech. Thus, not only were the creaky T4s very heterogenous in an uncontrolled way, but also, there was no way to control for absolute f0 or f0 movement in noncreaky regions of the naturalistic stimuli. Thus, we were unable to obtain direct information about sensitivity of listeners to details of creak, and we could not factor out creaky voice from other possible cues, such as low and/or falling pitch preceding the creaky region, or even the percept of low pitch in creak.

3.3 Experiment 2: Resynthesis of f0 and creaky voice quality

Experiment 1 showed that the presence of creak aided Cantonese listeners in identifying Tone 4 and discriminating Tone 4 from Tone 6. In this follow-up experiment, our goals were: (1) to attempt to demonstrate sensitivity to creaky voice independent of the concomitant low absolute pitch or pitch movement cues that may have been present in Experiment 1, (2) to probe the integration of creaky voice with contextual pitch cues for discriminating Tone 4 from Tone 6, and (3) to examine how/if the variation in the realization of creaky voice influences tone perception. For this purpose, we resynthesized and cross-spliced speech materials from the production corpus described in §3.2.1.1 to generate stimuli pitting pitch vs. creaky voice cues for a two-alternative forced choice task between lexical items with Cantonese Tones 4 and 6.

Using Wong and Diehl (2003)'s result that Cantonese listeners use the f0 of the preceding context to judge relative pitch for identifying a tone, we manipulated pitch perception by manipulating the f0 of the preceding context rather than the target syllable. Thus, the stimulus set consisted of disyllables, where the f0 of the first syllable was resynthesized to produce an 8-step continuum in increments of half-semitones, and the second syllable, the target syllable to be identified, was created by cross-splicing stimuli with different creak qualities. To follow up on Experiment 1 and try to demonstrate sensitivity to creaky voice independent of concomitant pitch cues, we also tested a monosyllable stimulus set consisting of only the target syllables.

To preclude the availability of absolute low pitch as a cue, we chose stimuli from a high-pitched male and female and resynthesized the f0 of the target syllable to be ambiguous between that of Tone 4 and 6 for the speakers, based on f0 ranges in the corpus and perception by two native speakers. Because the vocal fry mechanism is contingent on a low f0 (Gerratt and Kreiman, 2001), we selected instances of nonmodal phonation to cross-splice that were due to period doubling, with "pairs of vocal cycles alternating in period and/or amplitude" in which a pitch percept is ill-defined due to bitonality (Gerratt and Kreiman, 2001).

However, since a pitch in period doubled speech might still be detected based on the pulse width between glottal pulses, we included period doubled stimuli with varying pulse widths and strength of pitch percept. Finally, we varied the duration of the nonmodal region cross-spliced into the target syllable, since pitch was perceptible in the modal region of the target before the nonmodal region.

3.3.1 Methods

3.3.1.1 Materials

Productions of the Cantonese syllable /lau/ for Tones 4 and 6 were selected from one male and one female speaker with high pitch ranges out of the production corpus described in Experiment 1 §3.2.1.1. For each speaker, the utterance with the lowest level contour instance of *lau6* immediately following the sentence frame *lei5 jiu3* was selected. We did not want a f0 fall over the target syllable because that would introduce f0 information that might bias the listener towards a T4 response, and inspection of the production corpus indicated that both T4 and T6 occurred with level f0 variants. Three additional utterances were selected for each speaker to exhibit a range of variation in creaky realizations of Tone 4 immediately following *lei5 jiu3*. All Tone 4 selections were period doubled: one had a wider pulse width ("wide"), another a narrower width ("narrow"), and one had a very clear and audible pitch percept and its f0 was trackable by Praat's autocorrelation algorithm ("pitched") (Boersma and Weenink, 2010), see Figs. 3.1 and 3.2. For the male speaker, no "pitched" stimulus could be found, so a "pitched" stimulus from another male speaker was used. A stimulus was considered period doubled based on narrow-band spectrographic evidence of subharmonics.

The utterances were processed and resynthesized in Praat. The disyllable *jiu lau* was extracted for each utterance, and the f0 of the utterance was resynthesized using PSOLA as follows: (1) the absolute f0 of the diphthong /au/ was resynthesized to a value ambiguous between Tone 4/6 for high range males/females in the production corpus (180 Hz for the female; 107 Hz for the male); (2) the f0 of /jiu/ preceding /lau/ was resynthesized to be 31 Hz higher than the f0 of /au/, a relative f0 difference ambiguous between the f0 drop from T3 to T4/T6 in the production corpus for high pitch range speakers, and then incremented upwards and downwards in half-semitone steps from 1.5 semitones below to 2 semitones above that; (3) f0 was linearly interpolated in Hz over /l/ between the offset of /jiu/ and onset of /au/.

The creaky T4 /au/s were cross-spliced with the T6 in /jiu lau/ utterances. Durations of /jiu/, /l/ and /au/ were equalized to their averages between all utterances to facilitate naturalistic cross-splicing and also to standardize durations for reaction time measurements. The average amplitude of /jiu lau/ for the two non-creaky T6 utterances was resynthesized to 78 dB, and that of the /au/s extracted from the creaky T4 utterances to 72 dB, a lower value to aid in creating a continuous percept across the splice boundary, but high enough to not create a low amplitude cue for T4. For each creaky /au/ token, three splice points were chosen: "heavy", "medium", and "light", where "light" included the minimal amount of speech material to induce a creaky percept in the /au/ for the first author; "heavy" included the maximal amount of material that was creaky; "medium" was set at the approximate midpoint between the other two. All splice points were taken at the nearest zero crossing in a low amplitude regions. Monosyllable /lau/ stimuli were extracted from the /jiu lau/ stimuli.

The disyllable stimulus set included 3 repetitions of each modal stimulus, and 2 SPEAKER (male, female) × 8 CONTEXTUAL F0 SHIFT (8 f0 levels) × [3 CREAK TYPE (wide, narrow, pitched) × 3 CREAK PROPORTION (heavy, medium, light) +

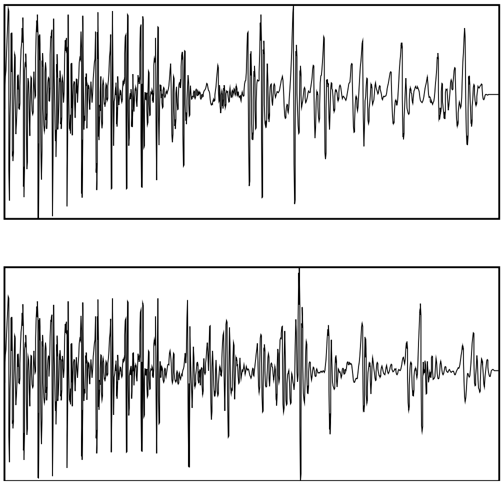


Figure 3.1: Waveforms of /au/ (17.85 ms) from the female narrow pulse width (top) and wide pulse width (bottom) bisyllabic stimuli with heavy creak proportion.

3 repetitions of modal stimuli] for 192 stimuli in total; the monosyllable stimulus set was balanced between creaky and modal stimuli: 2 Speaker × (3 creak type × 3 creak proportion + 9 repetitions of modal stimuli) for 36 stimuli in total.

3.3.1.2 Participants

The participants were 20 native Cantonese speakers who were born and raised in Hong Kong and currently living there. There were 10 males (age 20.3±1.9 years) and 10 females (age 21.8±1.7 years). They were recruited from the local Hong Kong university student population and received cash compensation.

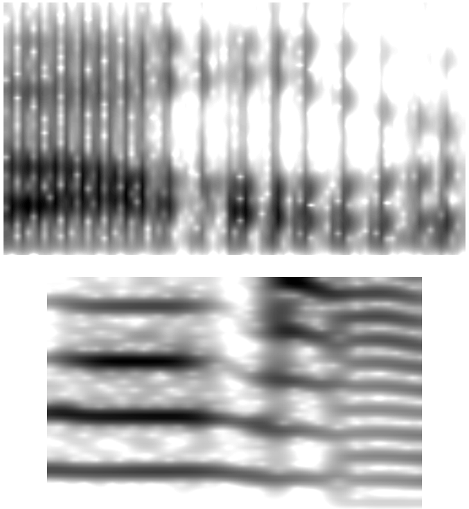


Figure 3.2: Spectrograms of the female bisyllabic stimuli. Wide-band spectrogram (max 5000 Hz) of /au/ from narrow width stimulus, showing doubled pulses (top). Narrow-band spectrogram (max 1000 Hz) of pitched stimulus, showing sub-harmonics (bottom).

3.3.1.3 Procedure

Participants were tested in a sound-treated booth. The perception experiment was run as in Experiment 1, except that the task of the participants was to identify each stimulus by a keyboard press of either a key labeled with the character for *lau4* (a common family name) or one labeled with *lau6* 'drip'. Participants were asked to respond as quickly and accurately as possible and told they would be timed. Each participant heard each stimulus set twice. The order of the different stimulus sets as well as which key was labeled with which word was randomized across participants, and participants received a short break between stimulus sets. Testing took about 30-40 minutes.

Participants were told that the stimuli were extracted from sentences *lei5 jiu3* __ *zi6* 'You want __ word' in the discourse context of looking up a word in a dictionary and a sheet with the sentence was placed before them during the experiment. Participants were also told that there was more than one speaker, that the speakers were asked to say the sentences in different pitch ranges, and that the relative proportions of the two different words played during each trial was randomized (so they wouldn't know what proportion to expect.).

3.3.1.4 Data analysis

The partial effects on response choice (T4 or T6) and log reaction time of: (i) CREAK (present,absent), (ii) details of creak, (iii) and for the bisyllables, also CONTEXTUAL F0 SHIFT, were analyzed using mixed effects regression analysis for the monosyllable and bisyllable stimulus sets using the procedures described in §3.3.1.4. Details of creak were expressed in two ways: (i) using two factors, CREAK TYPE (none, pitched, wide, narrow) and CREAK PROPORTION (none, light, medium, heavy), with the *none* levels dropped in analyses excluding the noncreaky stimuli, and (ii) using a factor crossing those two, CREAK QUALITY (10 levels: none, and the 3×3 crossing of the creaky levels of CREAK TYPE/PROPORTION). Following Wong and Diehl (2003, p. 47), CONTEXTUAL F0 SHIFT was treated as a continuous, interval-scale variable since it was based on the semitone scale.

REPLICATE was included as a noninteracted fixed effect covariate (since each stimulus set was presented twice). Unless otherwise indicated, the REPLICATE factor did not have a significant effect in model comparison and thus was omitted from final models. Because listeners showed systematically different patterns of behavior for the male and female stimulus sets, models were fitted to each of the two stimulus sets separately rather than including SPEAKER SEX as a covariate in a single model, except for one case, where the fixed effects structure was very simple (Table 3.5). Not including warranted random effects structure could result in anticonservative estimates of p-values for fixed effects (Janda et al., 2010, p. 43-44). Thus, in modeling random effects structure, we tested for the inclusion of random slopes and the correlation of random slopes with random intercepts (Baayen, 2008, p. 251-2), in addition to random intercepts, since exploratory data analysis suggested considerable individual variation not only in unconditioned T4 response bias and reaction times, but also in the effect of the variables of interest.

3.3.2 Results

Our general research questions for Experiment 2, reiterated, were: (i) *How do listeners integrate creaky voice with pitch cues?* and (ii) *Are listeners sensitive to details of creak?* To answer these, we analyzed both the lexical decision response choices (§3.3.2.1) and reaction times for the responses (§3.3.2.2).

In the analysis of response choice, before examining the interaction of the presence of creak and contextual f0 information, we first examined their independent effects—the effect of CREAK for the monosyllable stimuli (which had no immediate contextual f0 information) and the effect of contextual f0 information, CONTEXTUAL F0 SHIFT, for the noncreaky bisyllable stimuli. We asked: (a) *Does the presence of creak bias listeners toward T4 responses?* (b) *Does contextual f0 information bias responses, with higher f0 on the preceding syllable biasing for T4 responses?*, and then we followed up to address our main questions: (c) *How does the presence of creak interact with contextual f0 information for listeners?* (d) *Do listeners show sensitivity to glottal pulse width and duration of nonmodal phonation?*

3.3.2.1 Lexical decision responses

Does the presence of creak bias listeners toward T4 responses? Analysis of the monosyllable data indicated that the presence of creak biases for T4 responses in the absence of immediate contextual f0 information, as well as in the absence of absolute low pitch and pitch movement cues which may have been present in the naturalistic stimuli in Experiment 1. Likelihood ratio tests (Table 3.5) showed that the best model for the probability of a T4 response for the monosyllables included the following fixed effects: SPEAKER SEX, CREAK and the interaction SPEAKER SEX:CREAK and random slopes for CREAK correlated with intercepts for listeners, and uncorrelated random slopes for SPEAKER SEX. Thus, CREAK had a significant effect in determining the probability of a T4 response. The final model predictions for mean levels indicated that T4 responses are more likely for creaky stimuli than noncreaky stimuli by factors of 3.8 and 6.0 for the female and male stimuli, respectively. Inspection of individual subject data supported the random effects structure; there was variability in the effect on the probability of a T4 response across listeners due to the presence of creak, correlated with listeners' baseline bias for a T4 response, and between the male and female stimulus sets.

We also checked if the effect that the presence of creak biased listeners toward T4 responses held even for only stimuli with light CREAK PROPORTION. It did. Model comparison with random intercepts for listeners and correlated random slopes for CREAK within data subsets for only the light creak proportion vs. none contrasts supported the inclusion of CREAK in the model over one with no fixed effects (male: $\chi^2_m(1) = 17.56$, $p = 2.8 \times 10^{-5}$, female: $\chi^2_f(1) = 6.25$, $p = 1.2 \times 10^{-2}$). Because the narrow creak type male stimuli had particularly long intervals of nonmodal phonation, we also further checked that the T4 biasing effect of nonmodal phonation held for the non-narrow creak types for the male stimuli, and it did ($\chi^2(1) = 9.97$, $p = 1.6 \times 10^{-3}$).

Fixed effects included in model			
SEX	CREAK	SEX:CREAK	LL ratio test
✓			
✓	✓		$\chi^2(1) = 22.10, p = 2.6 \times 10^{-6}$
✓	✓	✓	$\chi^2(1) = 9.50, p = 2.1 \times 10^{-3}$

Table 3.5: Summary of comparison of mixed effects logistic models for T4 responses for monosyllable stimuli, including the fixed effects that were compared and the results of log-likelihood ratio tests for each successive comparison.

Does contextual f0 information bias responses? Analysis of the noncreaky bisyllable data indicated that contextual f0 information (CONTEXTUAL F0 SHIFT) biases responses: in the logistic model, the probability of a T4 response increases

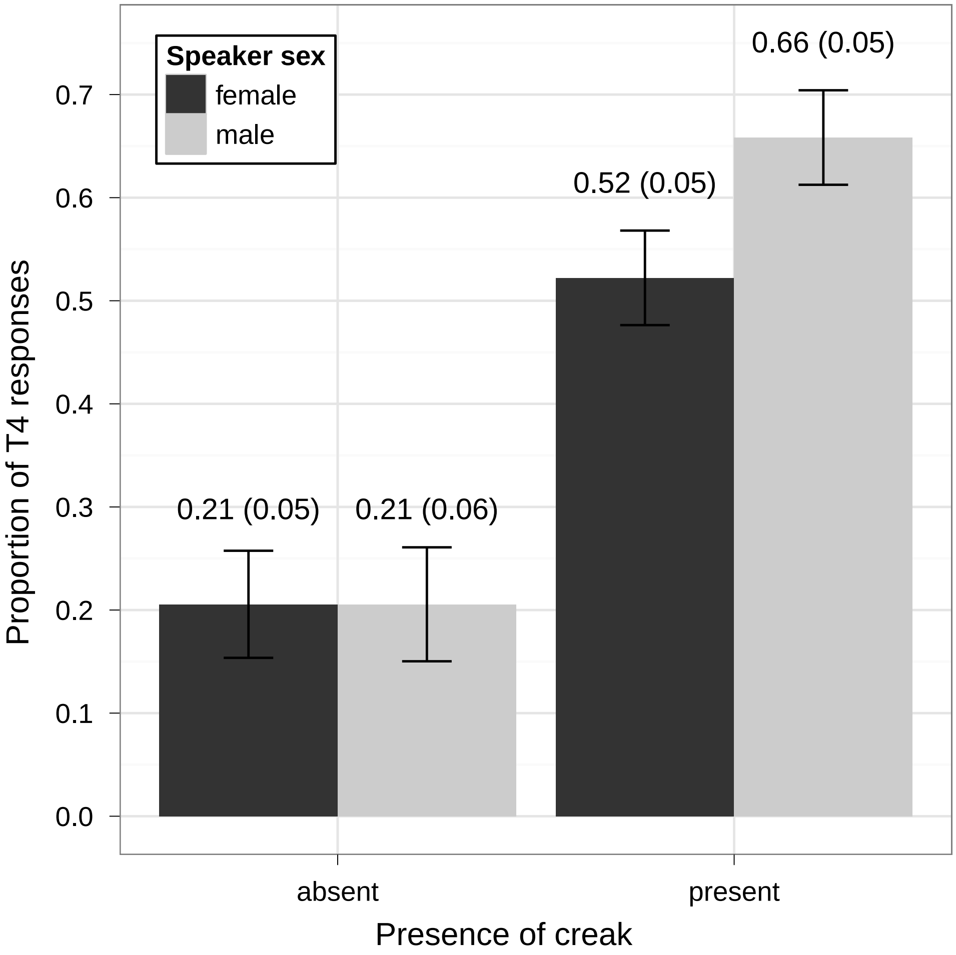


Figure 3.3: Overall proportion of T4 responses conditioned on presence of creak and stimuli speaker sex for monosyllabic stimuli. Error bars show ±1SE).

as f0 increases on the preceding syllable to the target syllable; we expected this since as f0 on the preceding syllable becomes higher relative to the target syllable, pitch on the target syllable is perceived to be contextually lower. For both the male and female stimulus subsets, model comparison indicated final models with random intercepts by-listener and uncorrelated random slopes for CONTEXTUAL F0 SHIFT. REPLICATE was also a significant covariate for the male subset, with a lower probability of T4 response in the second replicate. Likelihood ratio tests supported the inclusion of CONTEXTUAL F0 SHIFT as a fixed effect above the random effects (and REPLICATE covariate for the male stimuli), $\chi^2_m(1) = 20.21, p = 6.9 \times 10^{-6}$ and $\chi^2_f(1) = 31.38, p = 2.1 \times 10^{-8}$ for the male and female stimuli, respectively. The model coefficient for CONTEXTUAL F0 SHIFT for both stimulus sets was positive (Tables 3.6, 3.7), indicating that the probability of a T4 response increases with CONTEXTUAL F0 SHIFT in the noncreaky bisyllabic stimuli, consistent with the positive slope of the dashed-line T4 response curves for the noncreaky stimuli in Figure 3.4.

	Coefficient	SE	Z	*p*
(Intercept)	−1.77	0.39	−4.5	<.0001
REPLICATE	−0.64	0.19	−3.3	<.001
CONTEXTUAL F0 SHIFT	1.19	0.21	5.6	<.0001

Table 3.6: Summary of fixed effects for mixed logit model of T4 responses in noncreaky male bisyllables

How does the presence of creak interact with contextual f0 information for listeners? Model comparison for T4 responses in the full male and female bisyllabic stimuli supported the inclusion of both CONTEXTUAL F0 SHIFT and

	Coefficient	SE	Z	p
(Intercept)	−0.39	0.28	−1.4	>0.2
CONTEXTUAL F0 SHIFT	1.86	0.23	8.3	<.0001

Table 3.7: Summary of fixed effects for mixed logit model of T4 responses in noncreaky female bisyllables

CREAK as well as their interaction as fixed effects (Table 3.8). For both the male and female stimuli, the final models had positive coefficients for the main effects of CONTEXTUAL F0 SHIFT and CREAK but negative ones for the interaction (Tables 3.9, 3.10). That is, for both creaky and noncreaky stimuli, there is a higher probability of a T4 response with higher CONTEXTUAL F0 SHIFT, but for a given CONTEXTUAL F0 SHIFT, the probability of a T4 response is higher if the stimulus is creaky, and for the creaky stimuli, the increase in probability with CONTEXTUAL F0 SHIFT is smaller than for the is smaller than for the is smaller than for the noncreaky stimuli: the slope of the logit T4 response curve as a function of CONTEXTUAL F0 SHIFT is less steep for creaky than noncreaky stimuli. This is reflected in Figure 3.4, in which the response curves for the creaky stimuli are both globally shifted upward from and flatter than those for the noncreaky stimuli.

The final models included random intercepts by listener and random slopes for CONTEXTUAL F0 SHIFT and CREAK, with a correlated slope for CONTEXTUAL F0 SHIFT for the male stimuli. Inspection of individual listener data for the male stimuli supported the correlated random slope: listeners with steep slopes for the T4 response curve as a function of CONTEXTUAL F0 SHIFT, had a low proportion of T4 responses for low CONTEXTUAL F0 SHIFT, i.e., a low intercept, while subjects with flatter response curves had higher intercepts; for female stimuli, most subjects showed steep slopes and low intercepts.

Fixed effects included in model			
F0 SHIFT	CREAK	F0 SHIFT:CREAK	LL ratio test
✓			$\chi^2_m(1) = 14.10, p = 1.7 \times 10^{-4}$
✓	✓		$\chi^2_m(1) = 29.73, p = 5.0 \times 10^{-8}$
✓	✓	✓	$\chi^2_m(1) = 15.98, p = 6.4 \times 10^{-5}$
✓			$\chi^2_f(1) = 34.35, p = 4.6 \times 10^{-9}$
✓	✓		$\chi^2_f(1) = 18.45, p = 1.7 \times 10^{-5}$
✓	✓	✓	$\chi^2_f(1) = 26.72, p = 2.4 \times 10^{-7}$

Table 3.8: Summary of comparison of mixed effects logistic models for T4 responses for bisyllabic stimuli, including the fixed effects that were compared and the results of log-likelihood ratio tests for each successive comparison. Results for the male stimuli are shown, and then results for the female stimuli.

	Coefficient	SE	Z	p
(Intercept)	0.19	0.22	0.9	>0.4
CONTEXTUAL F0 SHIFT	0.68	0.14	4.8	<.0001
CREAK	2.30	0.29	7.8	<.0001
CONTEXTUAL F0 SHIFT:CREAK	−0.41	0.11	−3.9	<.0001

Table 3.9: Summary of fixed effects for mixed effects logistic model of T4 responses in male bisyllables

Do listeners show sensitivity to glottal pulse width and duration of nonmodal phonation? Model comparison for the creaky male and female monosyllabic and bisyllabic stimuli provided evidence for the inclusion of CREAK

	Coefficient	SE	Z	p
(Intercept)	0.84	0.19	4.4	<.0001
CONTEXTUAL F0 SHIFT	1.26	0.13	9.4	<.0001
CREAK	1.56	0.31	5.1	<.0001
CONTEXTUAL F0 SHIFT:CREAK	−0.56	0.11	−5.0	<.0001

Table 3.10: Summary of fixed effects for mixed effects logistic model of T4 responses in female bisyllables

QUALITY in models of the probability of a T4 response. We used treatment contrasts for CREAK QUALITY, without centering, as the data were balanced for this factor, and since we were simply initially checking for an effect of variability in the realization of nonmodal phonation. For the monosyllabic creaky stimuli, model comparison supported the inclusion of CREAK QUALITY in the model over one with only random intercepts for listeners (male: $\chi^2_m(8) = 131.25$, $p < 2.2 \times 10^{-16}$, female: $\chi^2_f(8) = 59.02, p = 7.24 \times 10^{-10}$). For the bisyllabic creaky stimuli, the inclusion of CREAK QUALITY was supported over models with only CONTEXTUAL F0 SHIFT and correlated random slopes for CONTEXTUAL F0 SHIFT and random intercepts by listener (male: $\chi^2_m(8) = 536.06, p < 2.2 \times 10^{-16}$, female: $\chi^2_f(8) = 220.97, p = 2.2 \times 10^{-16}$).

For a more detailed understanding of how details of the creak affected tonal perception, we performed model comparisons for the monosyllables with CREAK PROPORTION, CREAK TYPE, and their interaction. We parametrized the CREAK PROPORTION contrasts within the creaky stimuli as successive differences of the three levels: light, medium, and heavy (Venables and Ripley, 2002, p. 148-149). For both male and female stimuli, a higher proportion of creak significantly increased the probability of a T4 response (male: $\chi^2_m(2) = 32.31, p = 9.6 \times 10^{-8}$, female: $\chi^2_f(2) = 49.60, p = 1.7 \times 10^{-11}$) between both pairs of successive levels for the male stimuli, and between the light and medium levels for the female stimuli. Results for CREAK TYPE were inconsistent; for the male stimuli, the probability of a T4 response was highest for the narrow condition, followed by the wide and then the pitched condition, without any interaction with CREAK PROPORTION (Table 3.11); for the female stimuli, there was a significant effect of CREAK TYPE: CREAK PROPORTION, and the probability of a T4 response was higher for the wide pulse widths than narrow and pitched pulse widths, but only within the heavy proportion condition (Fig. 3.7). Within the heavy creaky female stimuli, a contrast for wide pulse width was supported by model comparison over one with only by-listener random intercepts: $\chi^2(1) = 7.60, p = 5.8 \times 10^{-3}$.

	Coefficient	SE	Z	p
(Intercept)	1.15	0.36	3.2	<.01
CREAK PROP.:MEDIUM-LIGHT	1.51	0.36	4.2	<.0001
CREAK PROP.:HEAVY-MEDIUM	0.96	0.39	2.5	<.05
CREAK TYPE:NARROW	3.47	0.43	8.0	<.0001
CREAK TYPE:WIDE	2.40	0.37	6.5	<.0001

Table 3.11: Summary of mixed logit model of T4 responses for CREAK PROPORTION and CREAK TYPE in creaky male monosyllables

For the bisyllabic stimuli, interactions between the creak quality factors were quite complex and we only present a sketch of patterns of results based on exploratory data analysis here, as the main result that details of the nonmodal phonation affect listeners' tonal identification responses has already been shown from model comparisons with CREAK QUALITY. The main patterns of results are shown in Figures 3.6 and 3.7. The female and the narrow and wide CREAK

TYPE conditions for the male stimuli show that when CONTEXTUAL F0 SHIFT was at its upper limits so that the preceding syllable f0 was relatively high, the proportion of T4 responses tended to asymptote to a ceiling value, regardless of creak quality.

For the female stimulus set, there is a noticeable split between the light and no creak conditions vs. the medium/heavy creak conditions at the bottom of the range of CONTEXTUAL F0 SHIFT such that the light/no creak conditions yielded fewer T4 responses. There is also a split at the bottom of the range for CONTEXTUAL F0 SHIFT for the male stimuli, but in which the no creak/pitched stimuli and wide pulse width/light proportion stimuli yield few T4 responses, but the other wide pulse width and all the narrow pulse width conditions yield a large proportion of T4 responses. In fact, the proportion of T4 responses for the heavy narrow/wide conditions for the male stimuli is quite stable over CONTEXTUAL F0 SHIFT and appears to be at ceiling, around 80-90%; in contrast, the pitched stimuli show a much steeper slope, with the proportion of T4 responses increasing as CONTEXTUAL F0 SHIFT increases. At the very highest CONTEXTUAL F0 SHIFT steps, the proportion of T4 response actually decreased for some listeners, which may have been due to stimuli naturalness problems.

3.3.2.2 Reaction times

Another measure of potential listener sensitivity to creak and details of creak is how the presence of creak and creak quality affected (log-transformed) reaction times for listener response. To probe this, we analyzed the effect of the presence of creak on log-transformed reaction times for T4 responses for the male and female monosyllables and bisyllables, and then also the effect of CREAK TYPE and CREAK PROPORTION within creaky subsets of those stimuli. In the analyses, we excluded subjects from the analysis who did not have T4 responses for all conditions for the factor being tested. For the bisyllables, we restricted the analysis to the reference level "0" CONTEXTUAL F0 SHIFT level, in which listener response was both close to a halfway split between T4 and T6 overall, and also all but one listener had T4 responses for each CREAK PROPORTION condition.

Model comparison with linear mixed models yielded no support for an effect of the presence of creak on log-transformed reaction times for any of the stimulus subsets. However, we did find limited support for an effect of CREAK PROPORTION and CREAK TYPE for T4 responses, within creaky stimuli. For the female creaky monosyllables, there was a trend for faster reaction times from medium to heavy creak proportion ($\chi^2(2) = 5.49$, $p = 0.064$ for the inclusion of CREAK PROPORTION in a null model with correlated random slopes for REPLICATE); for female creaky bisyllable stimuli at "0" CONTEXTUAL F0 SHIFT, model comparison supported the inclusion of CREAK PROPORTION for log-transformed reaction times for T4 responses ($\chi^2(2) = 10.68$, $p = 4.80 \times 10^{-3}$, comparing a model with CREAK PROPORTION to one without, with random slopes for REPLICATE). Reaction times were faster for heavy than medium creak proportion ($\beta_{medium-heavy} = -0.15$, $p_{MCMC} = 0.002$). At the highest CONTEXTUAL F0 SHIFT level, 2 semitones higher, we found no evidence for an effect of CREAK PROPORTION.

Finally we checked for an effect of CREAK TYPE in reaction times for the monosyllables. For the creaky male stimuli, with 5 subjects excluded, we found that reaction times were faster for the narrow pulse width condition relative to the other CREAK TYPE conditions ($\chi^2(1) = 4.94$, $p = 0.026$ for the inclusion of the fixed effect CREAK TYPE: NARROW) in addition to REPLICATE and random slopes correlated with by-subject intercepts for REPLICATE.

3.3.3 Discussion

In Experiment 2, we built on the result from Experiment 1 showing that listeners were sensitive to creak in Cantonese tonal perception with a more controlled experiment, where listeners were tasked to choose between T4 and its most confusable tone, T6. We controlled for f0 preceding the region of nonmodal phonation, creating an 8-step continuum of f0 on the syllable preceding the syllable to be identified, and interpolated from that f0 through the onset consonant to a constant f0 at the onset of the target vowel, in an f0 range ambiguous between T4 and T6. In the target vowel, we used cross-splicing to control the duration of the nonmodal region, as well as characteristics of the glottal pulse train in that region, and the nonmodal regions were period doubled, with a bitonal pitch percept.

Under these controlled conditions, we found, firstly, that evidence that the presence of creak biases listeners towards T4 responses, in the absence of immediate contextual f0 information, when only the isolated target syllables were presented, and in the absence of absolute low pitch and pitch movement cues that may have been present in the naturalistic T4 stimuli in Experiment 1. This bias was present even for the subset of creaky stimuli with a light proportion of creak, where creak was just perceptible.

From analyzing listener responses for the subset of bisyllabic stimuli that were noncreaky, we found that contextual f0 information alone also biases listener responses: the proportion of T4 responses increased as f0 on the preceding syllable increased. Wong and Diehl (2003) previously showed that preceding f0 strongly biases listener responses for the Cantonese level tones, T1, T3, and T6, and Huang and Holt (2009) showed that preceding f0 also affected perception of the Tone 2 rise in Mandarin, the first clear demonstration of an effect of preceding context on contour tone perception. As described in §3.3.1.1, we exploited allophonic variation which can cause T4 and T6 both to appear to have level f0 contours, although T4 is considered a fall, based on its citation form, while T6 is considered a level tone. Thus, we showed that preceding context affects contour tone perception in Cantonese, although we tested listeners on only level variants of T4 since we were controlling for pitch cues in the experiment.[8].

For the noncreaky bisyllabic stimuli, we also found that for the male stimuli, the proportion of T4 responses decreased for the second replicate; on inspection of individual listener data, we found that this pattern occurred for about 3 subjects, who showed about a 20% drop in T4 responses, whereas a handful of subjects exhibited the opposite pattern. Perhaps some subjects became sensitized to the presence of creak and this caused them to shift their decision rule for T4 to more heavily weight the presence of creak.

We also found that the presence of creak interacts with contextual f0 information for the listener: the presence of creak biased listeners towards T4 responses, but to different degrees depending on the f0 of the preceding syllable. When the f0 on the preceding syllable was at its lowest, the presence of creak could outweigh the preceding f0 information such that in the presence of creak at a given f0 shift, listeners sometimes changed their identification response from T6 to T4. When the f0 on the preceding syllable was at its highest for female stimuli, the presence of creak did not greatly affect tonal perception because the f0 information already biased listeners towards a T4 percept, which the presence of creak must have reinforced. Patterning along these lines, for the highest CONTEXTUAL F0 SHIFT level, reaction times for T4 responses for the creaky female bisyllable stim-

[8] While every noncreaky target syllable had a level f0 contour, if creak provides a kind of low pitch percept, then perhaps a stimulus with creak was effectively a contour tone, since its f0 contour fell into a creaky region persisting to the offset of the stimulus

uli were unaffected by creak proportion, although reaction times were faster for heavy than medium creak proportion for an f0 shift of 0, in the middle of the contextual fo shift range. Overall, the presence of creak had the effect of diminishing the effect of the preceding f0 context on tonal perception: the T4 response curves for creaky stimuli as a function of preceding f0 were flatter than those for the noncreaky stimuli. In the middle range for the f0 of the preceding syllable, the presence of creak appeared to be integrated with f0 information additively, on average across listeners, as the slopes of the response curves for the creaky and non-creaky stimuli were similar (Figure 3.4).

Finally, Experiment 2 provided evidence that listeners are sensitive to details of creak such as glottal pulse width and the duration of nonmodal phonation, since the inclusion of creak quality (crossing creak type and creak proportion) was supported by model comparison for both the creaky subsets for both the male and female monosyllabic and bisyllabic stimuli. Generally, there was a higher T4 response proportion as creak proportion increased, for male and female monosyllable and bisyllable stimuli. For the female bisyllable stimuli, the reaction times for T4 responses were also faster for heavy creak proportion than medium at the reference level "0" of contextual fo shift, as mentioned above, and medium and heavy creak proportion stimuli generally yielded higher proportions of T4 responses than the light creak proportion stimuli, particularly when the preceding syllable had low f0 (a T6 biasing context).

There were also effects of creak type. We expected, for instance, the wide glottal pulse stimuli to provide a relatively lower pitch percept than the narrow glottal pulse stimuli and thus favor a T4 response. Indeed, for the female monosyllables, the wide glottal pulse condition significantly increased the probability of T4 response relative to the narrow glottal pulse stimuli and stimuli with strong pitch percept, for the stimulus subset with heavy creak proportion. For the male monosyllables, though, the narrow pulse width stimuli actually yielded a higher T4 response proportion than the wide condition, but this may have because the narrow pulse condition for the male had longer durations of nonmodal phonation than the other creak type conditions; the narrow pulse condition also yielded faster reaction times than the other creak types. For the bisyllables, the effect of creak type was less clear. In the male bisyllables, the pitched stimuli generally yielded a relatively lower proportion of T4 responses especially when the preceding syllable had low f0 compared to the wide and narrow glottal pulse width stimuli, we are hesitant to take this as evidence for a strong effect of CREAK TYPE because the pitched condition for males involved cross-splicing speech produced by another vocal tract.

Overall, we did see some evidence for an effect of CREAK TYPE on tonal perception, which was sometimes consistent with listener sensitivity to pitch percepts in period doubled regions (for the female monosyllables), but the effect of duration of nonmodal phonation (CREAK PROPORTION) was more robust.

3.4 General discussion

In this paper, we showed that listeners are sensitive to the presence and detailed properties of creaky voice in native lexical tone perception in Cantonese, a language in which voice quality cues are considered to be non-contrastive in tonal representation. To our knowledge, Experiment 1 is the first demonstration that creaky voice can improve tonal identification accuracy in a tonal language with non-contrastive phonation. Previous experiments on the role of creaky voice in Mandarin tonal perception either could not show improvement in tonal identification accuracy because listeners were at ceiling in the task.

Both Experiment 1 and 2 demonstrated that voice quality cues affect tonal perception, even in languages without contrastive phonation. Thus, it is necessary to consider voice quality-related parameters for understanding human tonal representation in a potentially wide range of tonal languages. Furthermore, Experiment 2 showed that listener sensitivity to properties of creaky voice, the percept of a certain range of nonmodal phonation, affects linguistic cognition. Listeners' tonal identification and reaction times for identification were affected by not only the duration of nonmodal phonation in the speech signal, but also by characteristics of the glottal pulse train.

In all, our perceptual studies suggest that tone languages with non-contrastive phonation have a system of suprasegmental lexical contrast on the vowel/rime that is isomorphic to that of register languages: in these tone languages, while pitch cues are criterial for defining tonal categories, *tendencies* for phonation characteristics are criterial as well; in register languages, while phonation cues are criterial for defining registers, tendencies for pitch characteristics are criterial as well. Just as Henderson (1952, p. 151) writes of Cambodian (Khmer), that "the pitch ranges of the two registers may sometimes overlap, though what I shall call the second register tends to be accompanied by lower pitch than the first register", so too can creaky voice accompany both T4 and T6 (and all other tones) in Cantonese, though T4 tends to be realized with creaky voice more often. Furthermore, given more recent evidence that pitch cues may be criterial in register languages as well (Abramson et al., 2004, 2007), it seems that there may be a "fuzzy boundary" between tone and register languages (Abramson and Luangthongkum, 2009).

What our study is unable to directly address is how sensitivity to voice quality cues in tonal perception in a tonal language with noncontrastive phonation, as well as in tonal languages in general, is reflected in maps from the speech signal to tonal concepts. Similarly and more concretely, even if our results suggest that automatic tonal recognition would benefit by extracting acoustic parameters from the speech signal beyond f0, they do not pinpoint what these parameters might be.

Ideas about what these parameters might be can be fruitfully discussed in terms of automatic speech recognition, where the necessity of parameterizing the speech signal for computational modeling yields sharper definitions of speech sounds. Manual correction and smoothing of f0 contours or mapping missing values from f0 detection to a low f0 value below human pitch ranges to create a real-valued parameter cannot capture the results of Experiment 2. Listeners in Experiment 2 didn't simply categorize any stimuli with creak as T4, especially if the contextual pitch cues biased for T6 because of relatively low f0 on the preceding syllable. Thus, period doubled nonmodal phonation, at least, does not seem to simply be interpreted by listeners as an extra-low absolute f0 value in tonal perception. However, other parameters calculated as part of f0 detection algorithms are candidates for parameterizing the creaky voice percept for tonal recognition, such as correlation values from the generator function for candidate f0 estimates.

Work on the detection and classification of nonmodal phonation provides another source for potential parameters (Deshmukh et al., 2005; Surana and Slifka, 2006; Vishnubhotla and Espy-Wilson, 2007; Ishi et al., 2008) for the creaky voice percept in human tonal representation. Something common about parameter sets from this literature is that they all include parameters familiar from f0 detection. Surana and Slifka (2006) and Vishnubhotla and Espy-Wilson (2007) both use thresholds on pitch confidence, calculated from the autocorrelation peak and the average magnitude difference function, respectively, to classify frames with irregular phonation, and Ishi et al. (2008) uses a periodicity measure based on the autocorrelation function for the detection of vocal fry.

It may be methodologically useful to consider idealized f0 contours which are continuous, devoid of segmental and voice quality-related perturbations, but they seem to be a "rough and handy, seat-of-the-pants" abstraction rather than a "well-defined level of phonetic representation" (Pierrehumbert, 1990, p. 387). The similar parameters from f0 detection and the detection of irregular phonation underscore the findings of this paper that f0 is one of many interacting components of voice quality in representations of lexical tones in human cognition in a potentially wide range of tone languages, not just ones with contrastive phonation. This finding parallels recent perceptual work suggesting that neither phonation nor pitch cues alone, but both kinds of cues together, discriminate registers in human perception in Burmese, a register language (Gruber, 2011), and suggests more of a continuum between register and tone languages systems than sharp differences.

In fact, the effect of variable realizations of a phonation type on tonal perception illustrated here in a tone language with nonconstrastive phonation is likely to be present even in tone languages with contrastive phonation, although we know of no such studies addressing this. The lack of such studies is surely due in part to our poor understanding of how phonation types are generated, meaning that the knowledge necessary to design experiments finely controlling parameters of phonation is limited at best. Studying how pitch and phonation cues, their acoustic correlates, and other components of the voice source interact to understand human cognition and improve automatic tonal recognition can most efficiently proceed from interdisciplinary collaborations between engineers, psychophysicists, otolaryngologists, and linguists.

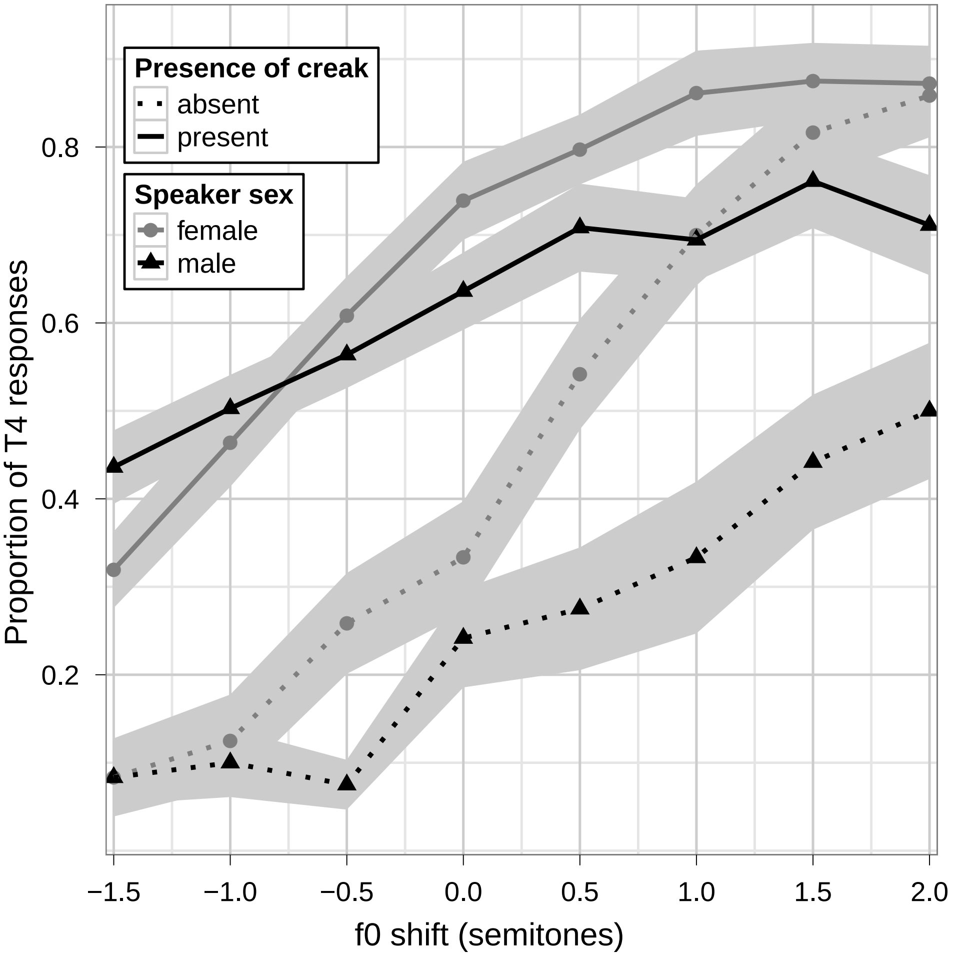


Figure 3.4: Overall proportion of T4 responses as a function of CONTEXTUAL F0 SHIFT conditioned on the presence of CREAK and SPEAKER SEX for bisyllabic stimuli, aggregated across listeners. The 0 point for CONTEXTUAL F0 SHIFT indicates the base resynthesized f0 level, from which the CONTEXTUAL F0 SHIFT continuum was created in increments of half-semitones. Ribbons show ±1SE. For the noncreaky stimuli, the response curve for female stimuli is much steeper than for male stimuli; for both the male and female stimuli, the response curve is less steep for the creaky stimuli and globally shifted upward from the response curve for the noncreaky stimuli.

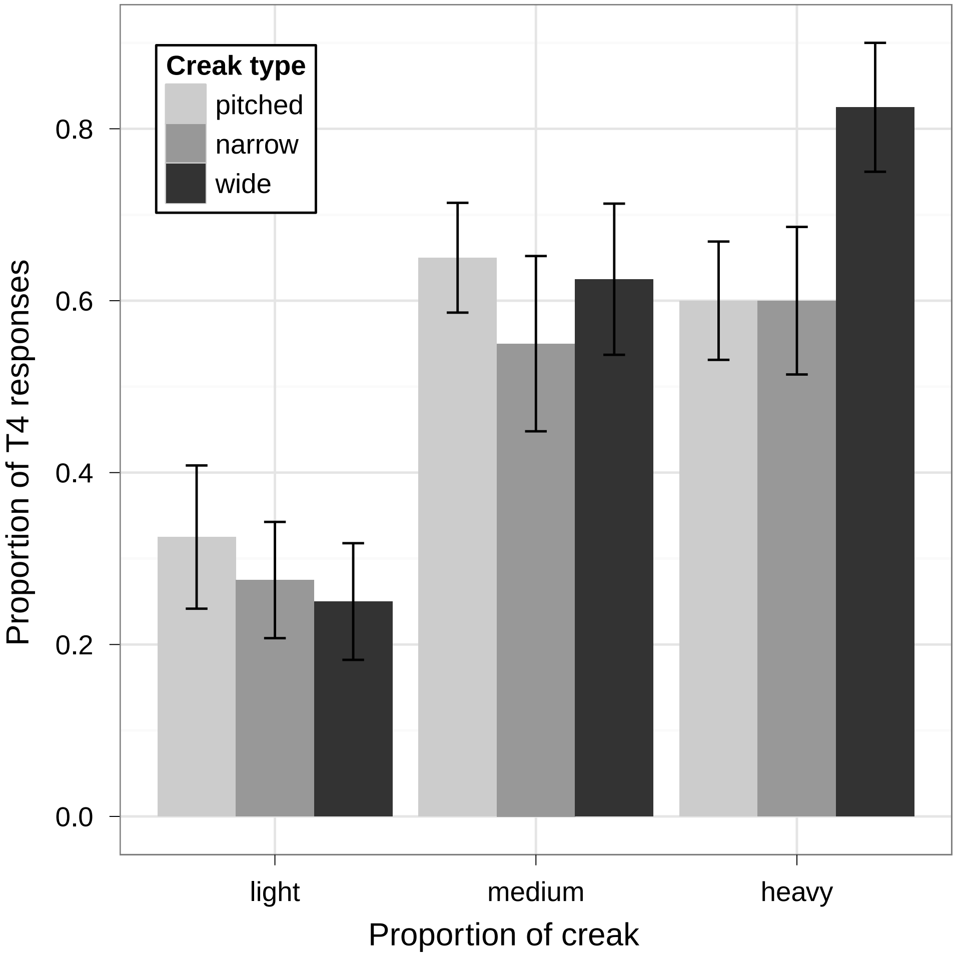


Figure 3.5: Overall proportion of T4 responses conditioned on the presence of CREAK PROPORTION within CREAK TYPE for female creaky monosyllabic stimuli, aggregated across listeners. Error bars show ±1SE. The proportion of T4 responses is higher for the wide creak type condition than the other two creak types within the heavy creak proportion condition.

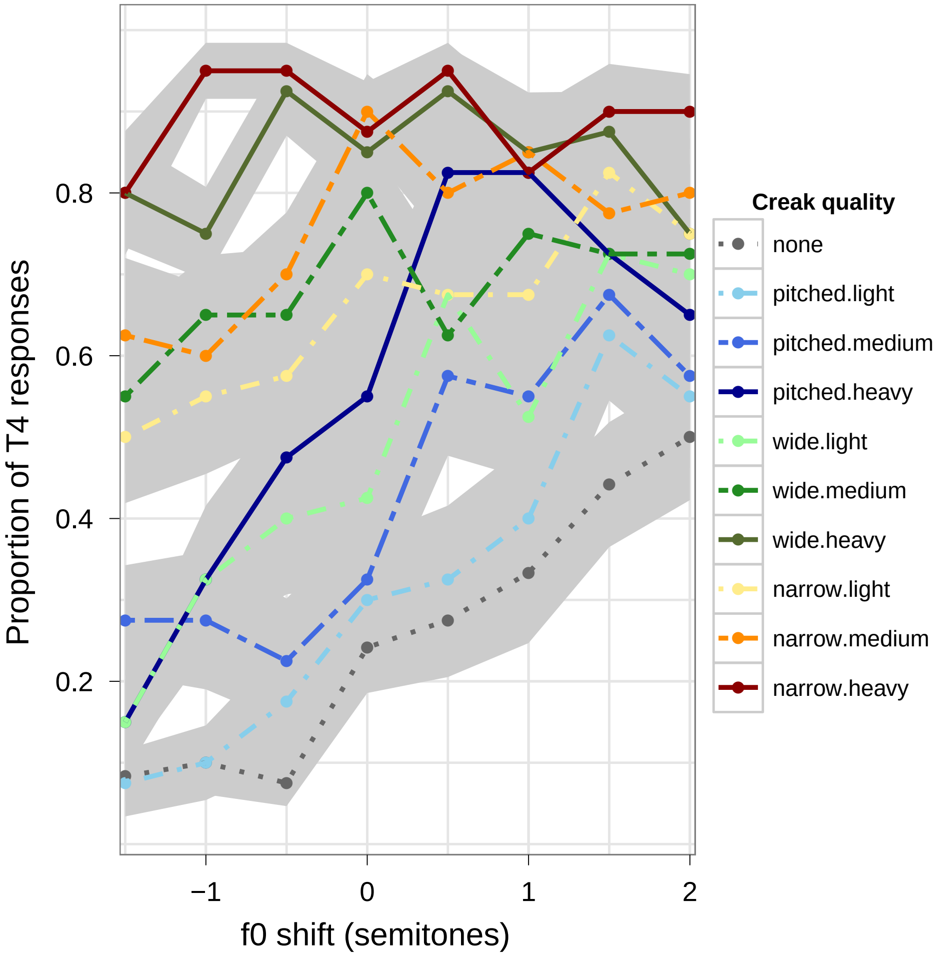


Figure 3.6: Overall proportion of T4 responses as a function of CONTEXTUAL F0 SHIFT, conditioned on CREAK QUALITY for male creaky monosyllabic stimuli, aggregated across listeners. Ribbons show ±1SE.

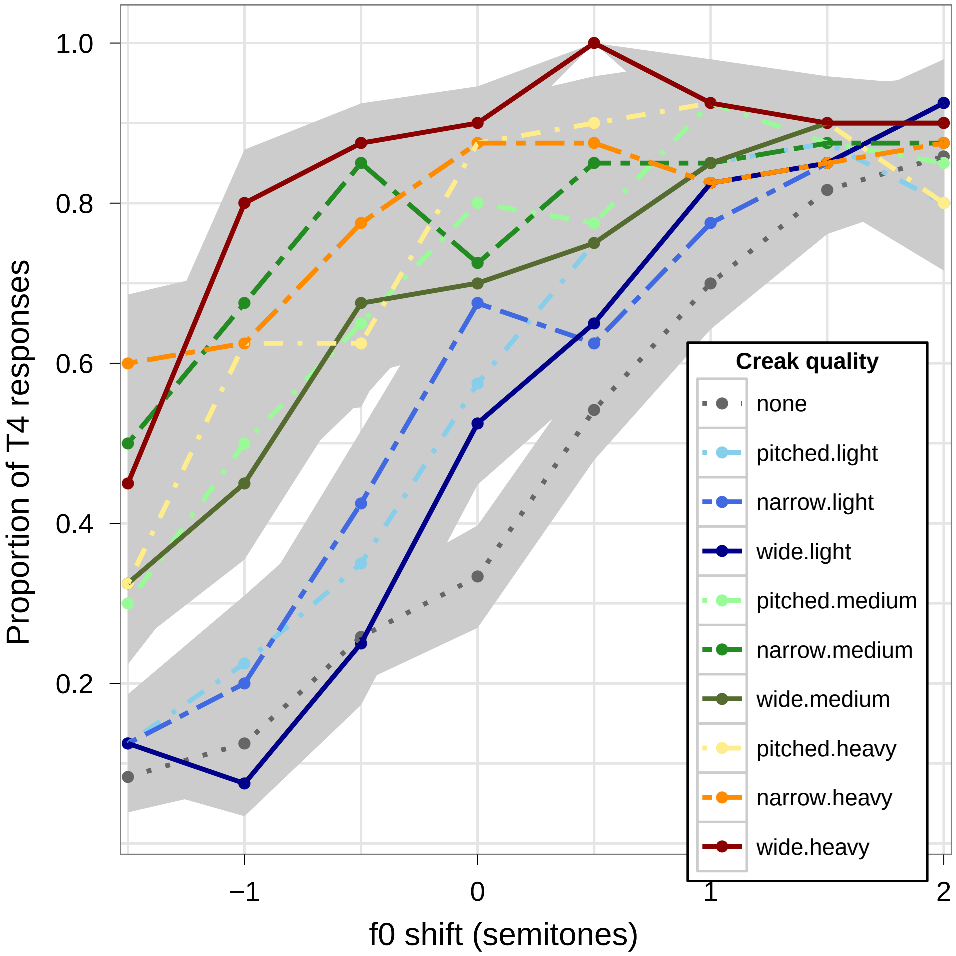


Figure 3.7: Overall proportion of T4 responses as a function of CONTEXTUAL F0 SHIFT, conditioned on CREAK QUALITY for female creaky monosyllabic stimuli, aggregated across listeners. Ribbons show ±1SE.

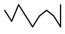
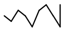
CHAPTER 4

Temporal resolution in tonal representations: a case study with Cantonese tones

4.1 Introduction

While the previous two chapters have addressed potential sources of complexity in the definition of tonal maps, this chapter considers evidence consistent with restrictive structure in the definition of tonal maps. The inclusion of contextual information (Chapter 2) and voice source information beyond fundamental frequency (Chapter 3) in the domain of tonal maps potentially contributes to complexity in the hypothesis space of possible tonal maps because it increases the dimensionality of tonal spaces.

However, viewed through the lens of Vapnik-Chervonenkis (VC) dimension, the cardinality of the set of parameters used in describing the class of possible tonal maps—the usual sense of dimensionality—is not the most revealing complexity metric for characterizing the learnability of the class of possible tonal map: a class of maps defined with just one parameter may have infinite VC dimension. A frequently given example of such a class is the family of indicator functions of sinusoids of arbitrary frequency $I(\sin(\alpha x))$ (Hastie et al., 2009, 237–238). As the frequency of the sinusoid, α, approaches infinity so does the number of points that can be shattered—this family is arbitrarily wiggly. In contrast, the family of indicator functions from rays defined as a set of real numbers greater than a threshold θ, $r(\theta) = \{x \in \mathbb{R} \,|\, \theta \leq x\}$—also defined with a single dimension—has a VC dimension of 2. The rays define a much more rigidly structured hypothesis space than the sinusoids do, though both define 1-dimensional learning problems.

Is the class of possible tonal maps arbitrarily wiggly? Considering for the sake of illustration tonal maps defined over only pitch values, are there tonal maps including mappings  → Tone X,  → Tone Y? The typological evidence available tentatively suggests not, at least for tones uttered in isolation. Maddieson (1977, 1978) make a distinction between "simple" and "complex" contours: a complex contour has an inflection point (it is minimally bidirectional), and while Maddieson (1978, p. 347) alludes to the occurrence of tridirectional contours,[1] Maddieson (1977, 1978) only explicitly discuss fall-rises and rise-falls as complex contours. Stating that there is a (small) finite bound on the wiggliness or flexibility of the class of tonal maps is an informal way of stating that the class has (small) finite VC dimension, which is precisely the characterization of the class that we discussed as guaranteeing learnability in Chapter 1. Thus, finding evidence consistent with such a bound is an important step in establishing the learnability of tonal maps.

One consequence of a constraint on the wiggliness of tonal maps would be that fine-grained attention to the unfolding in time of utterances of tones would not be necessary for good separability in the classification of tones of any tone language. For instance, if the class of tonal maps consisted of 2nd order polynomials defined over some space, then a minimum of two samples over the relevant tonal domain would be required to define the polynomial, but if the class was much more wiggly and consisted of 7th order polynomials, then the minimum number of samples

[1] Jun (2000) describes boundary tones in the Korean intonation system as having even more inflection points than this—up to five for the HLHL% or LHLH% tones.

would be seven.[2] Thus, the goal of this chapter is to establish evidence that *fine-grained attention to temporal resolution in the speech signal is not necessary for good separability in the classification of the tones of Cantonese.* This negative result would be admittedly weak but nevertheless positive evidence for the finite VC dimension of tonal maps in human languages.

Previous studies that have touched on the wiggliness of tonal maps have studied the use of piecewise linear or polynomial functions for approximating pitch contours, e.g. Hirst and Espesser (1993); Taylor (2000); Andruski and Costello (2004); Kochanski et al. (2005); Hermes (2006). Perceptual studies that present listeners with tones resynthesized with particular restrictive parameterizations, e.g. as polynomials of a certain degree, can probe if the restrictive parameterization chosen is a close approximation to that in actual human tonal maps by collecting similarity judgments between the resyntheses and original stimuli (Li and Lee, 2007). In this chapter, we back off from imposing a hypothesized restrictive parameterization on the listener and focus on manipulating the sampling resolution of the speech signal available to the listener by replacing uniformly spaced intervals of the speech signal with noise. This syntagmatic issue of sampling resolution is a more general one than the exact (or approximate) restrictive parameterization of tonal maps—it is about establishing the *existence* of such a restrictive parameterization—and it is orthogonal to questions about the paradigmatic set of parameters that are referred to in tonal maps (e.g. do tonal maps reference amplitude cues?): the listener may sample whatever information from

[2] A related idea from signal processing is *aliasing* in sampling, which is exploited in strobe light special effects on the dance floor in discotheques. A dancer under a strobe light appears to move in discrete steps at a slower speed than he is actually moving because the strobe light is flashing (sampling the movement) at a rate much slower than the dancer's movements. If a signal is undersampled, then the samples taken cannot distinguish between the original signal and an alias of lower frequency.

the speech signal is available.[3]

To our knowledge, this is the first perceptual study explicitly about sampling resolution in tones. It has been long been taken for granted but not empirically validated that fine sampling resolution of the speech signal is not necessary in tones in linguistics, and discussions of sampling resolution have only appeared in computational studies. Chao, who introduced the iconic tone letters (Chao, 1930) used in the International Phonetic Alphabet for representing linguistic tone, wrote: "the exact shape of the time-pitch curve, so far as I have observed, has never been a necessary distinctive feature, given the starting and ending points, or the turning point, if any, on the five-point scale" (Chao, 1968, 25), and tone letters are understood to have up to 3 samples, e.g. ˨˩˧. Additionally, Laniran (1992) argues for two targets per tone in Yoruba, a tone language with high, mid, and low level tones, and Barry and Blamey (2004) argues for acoustic Cantonese tonal spaces in $\mathbb{R}^2$ defined over onset and offset f0 values based on perceptual dimensions hypothesized from multidimensional scaling analyses of cross-linguistic tonal perception (Gandour and Harshman, 1978; Gandour, 1981, 1983).

In support of the linguistic intuition about the sufficiency of sparse temporal resolution for good tonal separability, Tian et al. (2004)'s automatic tonal recognition study of Mandarin found that sparse temporal resolution, with 4 samples/tone, can outperform fine-grained sampling with 1 sample/10 ms, concluding that "detailed information is useless for tone discrimination" (Tian et al., 2004, I-107). However, in a study of unsupervised learning of Mandarin tones, Gauthier et al. (2007) extracted 30 samples of f0 or 28 samples of f0 velocity

[3] However, the manipulation used is limited to fixing the samples from which all parameters are extracted to be the same and thus to fixing the same sampling resolution across all parameters sampled from the speech signal.

per syllable, a frame shift on the order of 10 ms, and Zhang and Hirose (2004)'s hidden Markov model based Mandarin tonal recognizer used a 10 ms frame shift, while other Mandarin and Cantonese tonal recognizers have used a simple time warping-like (time normalization) sampling scheme of 3-5 f0 averages or frame values over uniformly divided subsegments of (part of) the syllable (Peng and Wang, 2005; Qian et al., 2007; Wang and Levow, 2008; Zhou et al., 2008). In sum, the computational literature has not settled on the fineness of sampling resolution to use in features extraction from the speech signal for tones.

To connect our perceptual study of sampling resolution with the tonal recognizer literature, we provided an experimental context for tonal identification limited in a way to be similar to characteristics of feature extraction in automatic tonal recognition. We used tritone stimuli, as most recent automatic tonal recognizers use acoustic feature extraction from a temporal window extending beyond a single tone to its neighbors (Zhang and Hirose, 2000; Levow, 2005; Qian et al., 2007), and we used stimuli from multiple speakers like in the speaker-independent tonal recognition tasks in Peng and Wang (2005); Qian et al. (2007). We also resynthesized the syllable durations of the tritones to be fixed at their grand average to simulate the commonly employed preprocessing step of time normalization to the syllable.

Our experimental manipulation of temporal resolution in the signal used interrupting noise to create a 5-step gradient of sampling resolution (frame shifts) and make uniformly distributed "samples" or windows from the speech signal available to the listener, a very simple treatment designed to simulate the common uniformly sampled vector time series feature extraction procedure in automatic tonal recognition. To address our research question, we compared the tonal identification accuracy of listeners between the different sampling resolution conditions. We also performed a proof-of-concept machine classification experiment of the experimental stimuli to see if the tones were well-separated under some known acoustic parametrization of the speech signal extracted with low sampling resolution to gain insight into how good perceptual separability of the Cantonese tones might be possible under low sampling resolution.

We chose to perform the experiment in Cantonese, for the ease of finding a sample of speakers large enough for the experimental design and because the six tones of Cantonese comprise a good exemplar tone inventory in having level tones (high level T1, 55, ˥; mid level T3, 33, ˧; low level T6, 22, ˨), rising tones (high rising T2, 25,˨˥; low rising T5 23, ˨˧), and falling tones (T4, 21, ˨˩), cf. Figure 4.2 (Matthews and Yip, 1994).[4] While most tone perception experiments have been done in Mandarin, the Mandarin tonal inventory lacks level tone contrasts, but most tone languages have at least one level tone contrast (Maddieson, 1978).

In the rest of the chapter, we discuss the speech materials used in the perception experiment and computational modeling (§4.2), the perception experiment (§4.3), the computational modeling (§4.4) and conclude with a general discussion (§4.5).

4.2 Speech materials

4.2.1 Recordings

The stimuli were recorded by ten native Cantonese speakers, five of whose recordings were further processed for the rest of the study: these three males and two

[4] Descriptions vary in the exact 5-value integers assigned to the tones, but the exact integers used here are not of importance; we use these designations as mnemonic names for the tonal categories throughout the paper. Some descriptions also distinguish these tones from the shorter entering tones (high, mid, and low level) which occur in syllables with unreleased stop codas.

females were chosen to span a wide pitch range, cf. §4.2.3, Table 4.1, to provide a representative instance of the challenge of a multispeaker task. Four of the speakers were born and raised in Hong Kong and recorded in the phonetics lab sound-attenuated booth at the City University of Hong Kong. One was born and raised in Macau and recorded in the phonetics lab sound-attenuated booth at University of California, Los Angeles. They were recruited from the local university student population and received cash compensation. All speakers were recorded digitally at 22,050 Hz/16 bits with PCQuirerX (Scicon R&D, Inc.) or at 44.1kHz/16 bits with a digital recorder.

The stimuli were created from the tritone ⟨ *wai*˧, {*wai*˥, ˨˥, ˧, ˨˩, ˨˧, ˨} , *mat*˧ ⟩ (*wai*^{33} *wai mat*^{3}) extracted from sentences of the form: *lei*$^{25/35}$ *yiu*^{33} *wai*^{33} *wai mat*^{3} *deng/geng*^{33} '*you want Wai-Wai to clean the lamp/mirror*' with the target, the second /wai/, ranging over all six Cantonese tones. The lexical meanings of the orthographic characters we associated with tones 55, 25, 33, 21, 23, and 22 were, respectively, 'power', 'appoint', 'fear', 'surround', 'great', and 'stomach', and speakers were asked to treat /wai wai/ as a (nonce) proper name. The orthographic characters were chosen to be the most familiar ones for each tone by a native speaker. Each speaker actually recorded 5 fluent repetitions of sentences containing all 36 bitone combinations over /wai wai/ (with the sentences not used as stimuli for the perception experiment serving as fillers), from which we chose the last three repetitions of each Tone 33-Tone X bitone for the stimuli set for a total of 90 tritones, 18 from each speaker, 3 distinct repetitions per speaker per Tone 33-Tone X bitone.[5] A Cantonese native speaker trained in linguistics and phonetics checked that none of the speakers had tonal mergers and that the speakers uttered the tones correctly. No speakers produced Tone 55 with a 53

[5] In three cases, we chose another repetition than those listed above due to sound quality of the recording.

high fall contour, a variant more common in the past.

4.2.2 Resynthesis

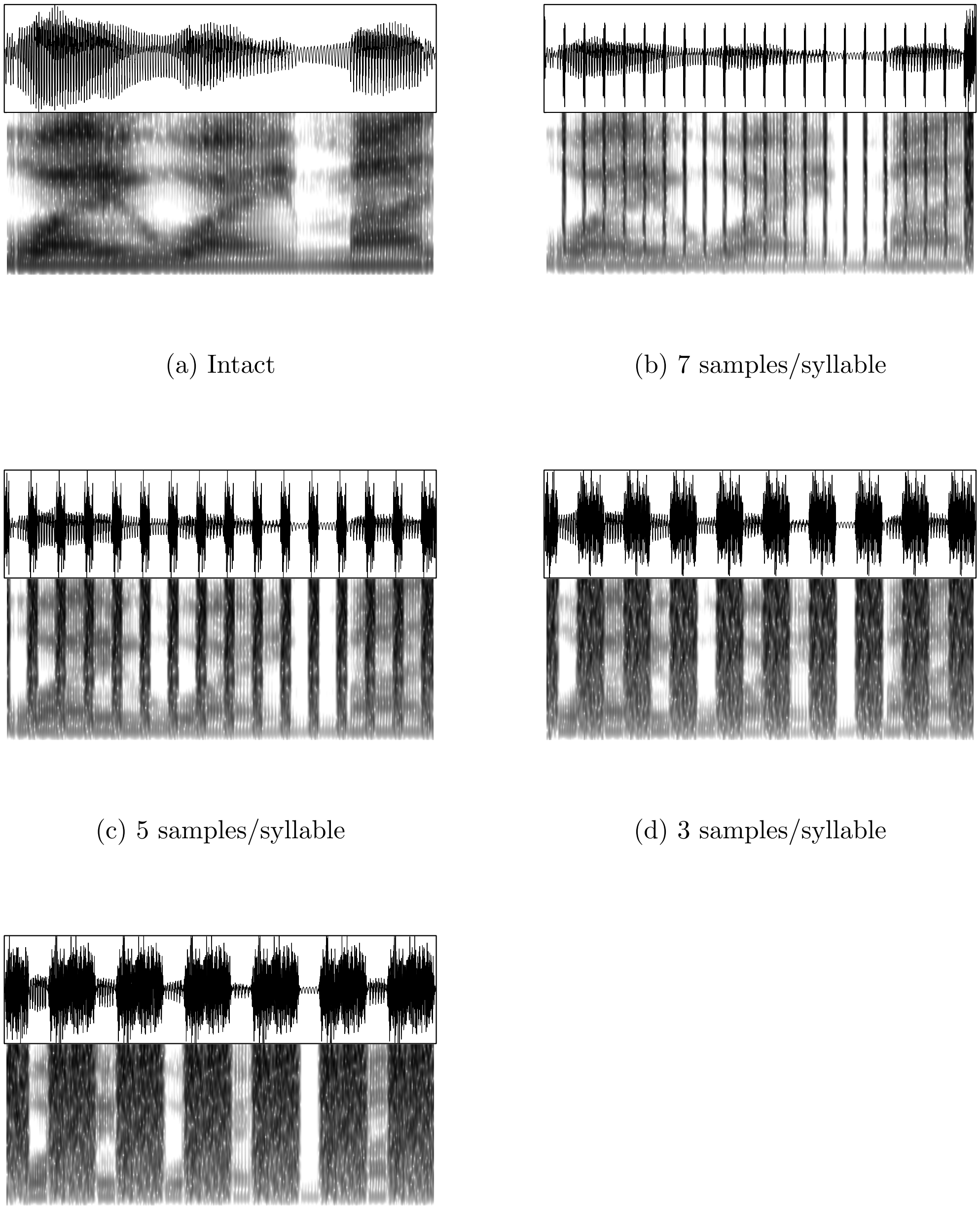
All stimuli were resampled to 22kHz; tritones were extracted using a rectangular window, and RMS amplitude was rescaled to 75 dB in Praat (Boersma and Weenink, 2010). All syllable durations were resynthesized using PSOLA implemented in Praat to be have a target duration of 241 ms, the grand mean of the syllable durations, for a total duration of 740 ms for the tritone, to simulate time normalization to the syllable.[6] The manipulated condition, SAMPLING RESOLUTION, was varied from the intact signal, to 7, 5, 3, and 2 uniformly spaced samples (time-slices or windows) of 30.41 ms each per syllable. The sample duration was well below the minimum 130 ms duration Greenberg and Zee (1977) found necessary for perception of a nonzero f0 velocity, "contouricity", in speech, and also on the same order of magnitude as the standard frame size in automated short-term analysis f0 detection (Hess, 1983, 343).

The sampling resolution manipulation involved intermittently deleting the recorded speech signal and replacing it with white noise low-pass-filtered at 5000 Hz that was 10dB higher than the average signal amplitude, cf. Figure 4.1. Similar stimuli manipulations are used in phonemic restoration studies, in which listeners perceive segmental speech sounds to be present in the presence of noise even if they are not (Miller and Licklider, 1950; Warren, 1970; Bashford et al., 1992; Samuel, 1996). We alternated the speech signal with louder noise rather than silence because the intelligibility of the speech is well-known to be minimal when alternated with silent gaps; however, continuity of the speech percept can be maintained when the speech signal is alternated with a louder sound that

[6] The PSOLA algorithm resynthesis added about 18 ms over the target duration over the course of the tritone.

is a potential masker of the fainter speech signal. This phenomenon is in fact the basis of phonemic restoration. Broadband noise has typically been used in segmental phoneme restoration experiments, and we chose to use white noise low-pass-filtered at 5000 Hz in particular because it has also been used in studying the continuity of tones through interrupting noise (Ciocca and Bregman, 1987). Additionally, we chose white noise to avoid providing any information that the listener might use in perceiving the interrupted speech, since Bashford et al. (1996) showed a boost in the intelligibility of speech interrupted by speech-modulated noise rather than white noise.

The noise was generated using using the MLP Matlab toolbox (Grassi and Soranzo, 2009). Since the sample durations were fixed, the noise duration varied for different SAMPLING RESOLUTION conditions, but was fixed within a condition, ranging from 90ms to 4ms from the 2- to 7- sample condition, respectively, as shown in Figure 4.1. The noise intervals included raised-cosine onset and offset ramps that were 10% of the duration of the noise interval to reduce audible spectral splatter (Hant and Alwan, 2003); the duration of the ramps was chosen to be relative to the duration of the noise interval since the noise interval duration varied between sampling resolution conditions. Half-duration noise intervals were used at the onset and offset of the tritone, with extra noise padding at the offset if needed to replace the entirety of the duration of the intact speech signal. Due to a programming error not detected until after the participants were tested, the last noise interval for the 2- and 3-sample stimuli was of full rather than half duration, extending beyond the duration of the intact tritone. However, the same information from the speech signal was available to the listeners that would have been present without the added noise.



(a) Intact

(b) 7 samples/syllable

(c) 5 samples/syllable

(d) 3 samples/syllable

(e) 2 samples/syllable

Figure 4.1: Waveforms and spectrograms of a Cantonese tritone Tone 33 - Tone 21 - Tone 33 stimulus under different sampling resolution conditions from intact, to 7, 5, 3, and 2 samples/syllable.

4.2.3 Acoustic analysis of resynthesized speech materials

We performed an acoustic analysis of the resynthesized speech materials based on extracted f0 tracks, cf. Fig. 4.2. The f0 values were extracted using RAPT (Talkin, 1995), a commonly used f0 detection algorithm, used in Qian et al. (2007)'s Cantonese supratone tonal recognizer. Speaker-specific pitch floors and ceilings were set to the 1st and 99th quantiles minus or plus 30% of the range, respectively, a similar procedure to the pre-processing procedures in De Looze and Rauzy (2009); Evanini and Lai (2010).[7] Otherwise, the default parameter settings, including a 10ms frame shift were used. The first and last frames were excluded because there were often large discontinuities between the estimated f0 for these frames and estimated f0 in the adjacent ones due to edge effects in the f0 detection algorithm, so there were a total of 69 f0 values, which were taken as the available f0 information in the intact condition. Unvoiced frames were assigned f0 values using linear interpolation. To model the f0 information present in the degraded SAMPLING RESOLUTION conditions, the mean f0 was calculated over each unmasked region over frames falling within each of these regions. Thus, there were 6 f0 values estimated for each tritone in the 2-sample condition, one per unmasked region, and 9, 15, and 21 f0 values in the 3, 5, and 7-sample conditions, respectively. The f0 values were also log-transformed and then standardized as z-scores using speaker-specific means and standard deviations. The calculated raw and transformed f0 range of the stimuli for each speaker is given in Table 4.1.

[7] The majority of the f0 values were in the mid range since each tritone stimulus consisted of two mid-level tones (33), yielding a center-heavy distribution of f0 values; thus, we could not use less extreme quantiles as in De Looze and Rauzy (2009); Evanini and Lai (2010) because they resulted in severe compression of the estimated range.

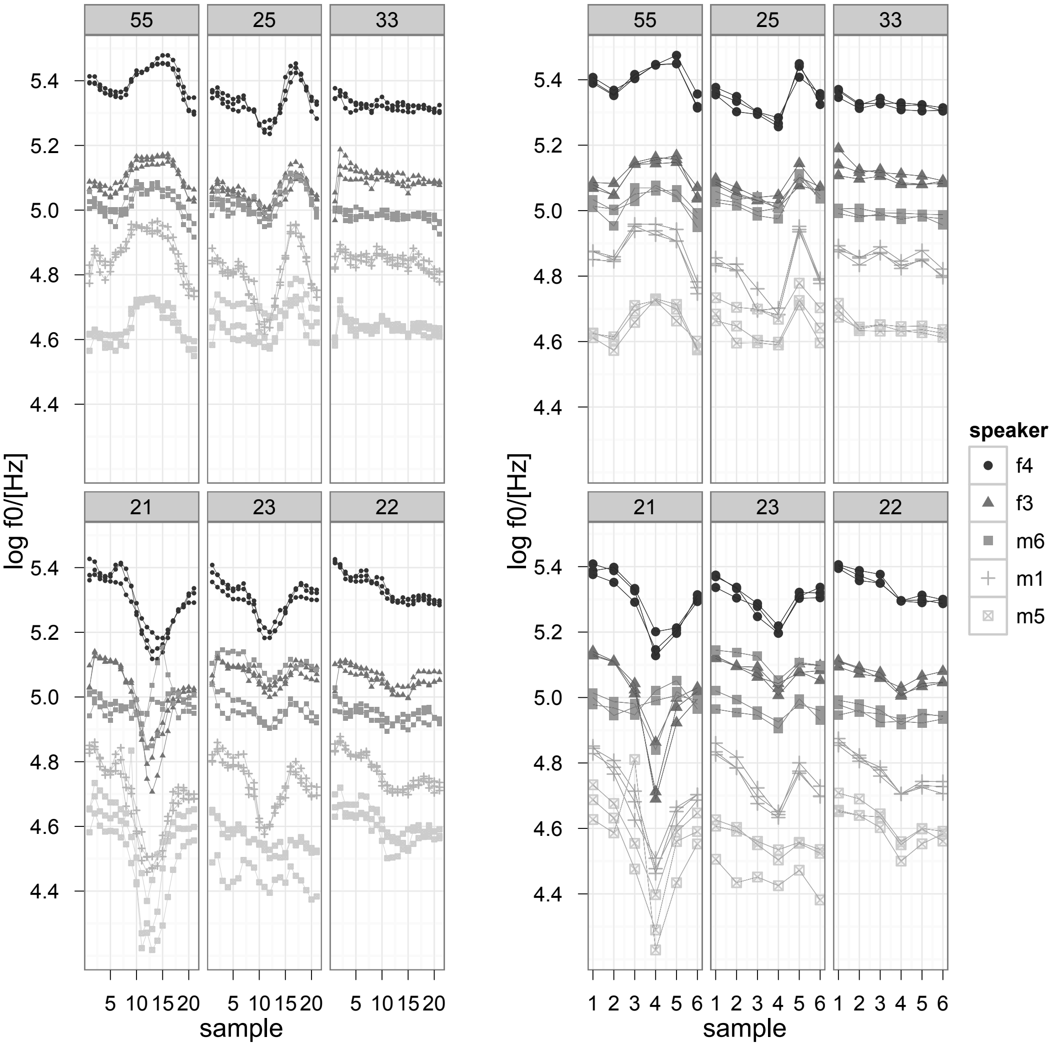


Figure 4.2: f0 contours extracted with RAPT using speaker-specific pitch floors and ceilings, showing the parameterization of f0 contours for the intact and 2-sample conditions for computational modeling. Unvoiced frames were assigned f0 values by linear interpolation. (left panel) log-transformed f0 extracted with 10ms frameshift and averaged over each of 21 samples in the 7 samples/syllable condition. (right panel) log-transformed f0 values averaged over each of 6 samples in the 2 samples/syllable condition.

Speaker	f0 (Hz)	log f0	z-score
f4	[165.89,241.00]	[5.11,5.48]	[-3.35,2.35]
f3	[106.42,179.47]	[4.67,5.19]	[-5.78,1.83]
m6	[125.88,176.36]	[4.84,5.17]	[-2.89,3.21]
m1	[83.87,145.92]	[4.43,4.98]	[-3.48,1.97]
m5	[61.44,140.20]	[4.12,4.94]	[-5.08,3.60]

Table 4.1: Speaker-specific f0 range in speech materials, measured in Hz, after log-transformation, and after standardization of log f0 with respect to speaker means and standard deviations. The speakers are ordered from highest to lowest maximum f0, following the same order from top to bottom in the plot of f0 contours by speaker in Figure 4.2.

4.3 Tonal perception experiment

Using the speech materials described in the preceding section, §4.2, we performed a human tonal perception experiment with native Cantonese speakers.

4.3.1 Methods

4.3.1.1 Participants

The participants were 39 native Cantonese speakers. There were 20 males (age 18.9±1.8 years) and 19 females (age 21.9±1.9 years). Participants were recruited from the local university student population at the City University of Hong Kong and at the University of California, Los Angeles and received cash compensation. All but three of the subjects (born/raised in Guangzhou and Shanwei, China) was born and/or raised in Hong Kong, China. Of the 10 participants tested in

Los Angeles, all used Cantonese on a daily basis and had been in the United States for 3 to 8 years.

4.3.1.2 Procedure

Participants were tested in sound-attenuated booths in the phonetics laboratories at the City University of Hong Kong and University of California, Los Angeles. The perception experiment was run in MATLAB using Psychophysics Toolbox extensions (Pelli, 1997; Brainard, 1997). Stimuli were played from an Echo Indigo IO sound card on a laptop over studio monitor headphones at a standardized, comfortable volume. The interstimulus interval was 3s.

Participants were told that the stimuli were extracted from sentences *lei*$^{25/35}$ *yiu*^{33} *wai*^{33} *wai mat*^{3} *geng*^{33} 'You want NAME to clean the mirror,' and they were given a sheet of paper with orthographic characters which showed what stimuli was being played, and what word they were to identify: *wai*^{33} __ *mat*^{3}. The stimuli were blocked by sampling resolution; block order was pseudorandomized to be roughly uniformly distributed over sampling resolution condition across participants, and stimuli were randomized within blocks. The task of the participants was to lexically identify the target syllable in each stimulus by a keyboard press of one of six keys labeled with the characters for the minimal tone set over *wai*. Participants were asked to respond as quickly and accurately as possible and told that they would be timed.

4.3.1.3 Data analysis

Statistical analysis was performed in R (R Development Core Team, 2010), and the ggplot2 package was used for creating graphics (Wickham, 2009). Tonal identification accuracy was analyzed using mixed effects linear regression (Pinheiro and Bates, 2000) implemented by the lme4 package (Bates and Maechler, 2010). The interest in this study, as in most psychological experiments, was to generalize beyond the sample of listeners and the sample of speakers from which stimuli were drawn. Mixed effects models allowed the inclusion of both the listener and the speaker as crossed random effects (Baayen et al., 2008), allowing simultaneous generalization to other listeners and speakers (Quené and van den Bergh, 2008).

Forward model selection was used to test the partial effects of SAMPLING RESOLUTION and the inclusion of random effects in modeling tonal identification accuracy aggregated over tones and for each individual tone. Successive nested models were compared using likelihood ratio tests, and χ^2 tests were used to test for significant improvement in fit to the data while penalizing for model complexity (Baayen, 2008, p. 253), since differences in deviance ($-2\log(\mathit{likelihood})$) between nested models fit to the same data by maximum likelihood approximately follow a χ^2 distribution in the large-sample limit. All models of identification accuracy included random intercepts by listener and speaker (of the stimuli), and all models except that for Tone 21 identification accuracy also included random slopes by listener correlated with the random intercepts by listener for SAMPLING RESOLUTION, since model comparison did not support the inclusion of the random slopes for Tone 21. The inclusion of listener-specific random effects in the models helped account for listener variation in tonal identification accuracy in the intact condition and the effect of SAMPLING RESOLUTION on accuracy, as well as for the effect of block order variation between listeners.

Tukey tests for all pairwise comparisons of SAMPLING RESOLUTION conditions were performed on the mixed effects models using the multcomp package (Hothorn et al., 2008) to compare tonal identification accuracy between SAMPLING RESOLUTION conditions. The Tukey tests were based on t distributions with 22 degrees of freedom: 4 degrees of freedom for the 5-level fixed effect SAMPLING RESOLUTION, 2 in total for the variance of each of the two random intercepts, 4 for the variance parameters for the random slopes for SAMPLING RESOLUTION, 10 for the correlation parameters for the random slopes ($\binom{5}{2}$ since the fixed effect had 5 levels), 1 for the fixed effect intercept term, and 1 for the residual variance; for the model of Tone 21 accuracy which didn't include random slopes, t distributions with 8 degrees of freedom were used. Statistical significance wherever discussed was determined at the 0.05 level.

4.3.2 Results

In the following two sections, we report on overall tonal identification accuracy (§4.3.2.1) and identification accuracy of individual tones (§4.3.2.2) conditioned on sampling resolution. All listeners and items were included in the analyses. While all listeners performed at above chance levels in the intact condition overall, not all listeners performed above chance levels for each individual tone, cf. §4.3.3, and in addition, there were three items that were not identified at above chance levels. However, when we repeated statistical analyses excluding these listeners and items, the statistical pattern of results did not change.

4.3.2.1 Overall tonal identification accuracy conditioned on sampling resolution

As Figure 4.3 shows, tonal identification accuracy aggregated across listeners was well above chance for all sampling resolution conditions, even down to 2 samples/syllable, when there was less than a quarter of the original speech signal present. Bonferroni-corrected t-tests of by-subject tonal identification accuracy against the at-chance level (1/6) showed that performance for each condition was significantly above chance, $p < 2.2 \times 10^{-16}$ for every condition. In addition, mixed effects linear regression indicated that aggregated across tones, tonal identification accuracy for the 5 and 7 samples/syllable conditions was not significantly different from that in the intact condition. Since exploratory data analysis showed that the effect of SAMPLING RESOLUTION on tonal identification accuracy varied by tone, though (see Figures 4.4 and 4.5), we focus our description of statistical results on separate models of identification accuracy for each of the six tones.

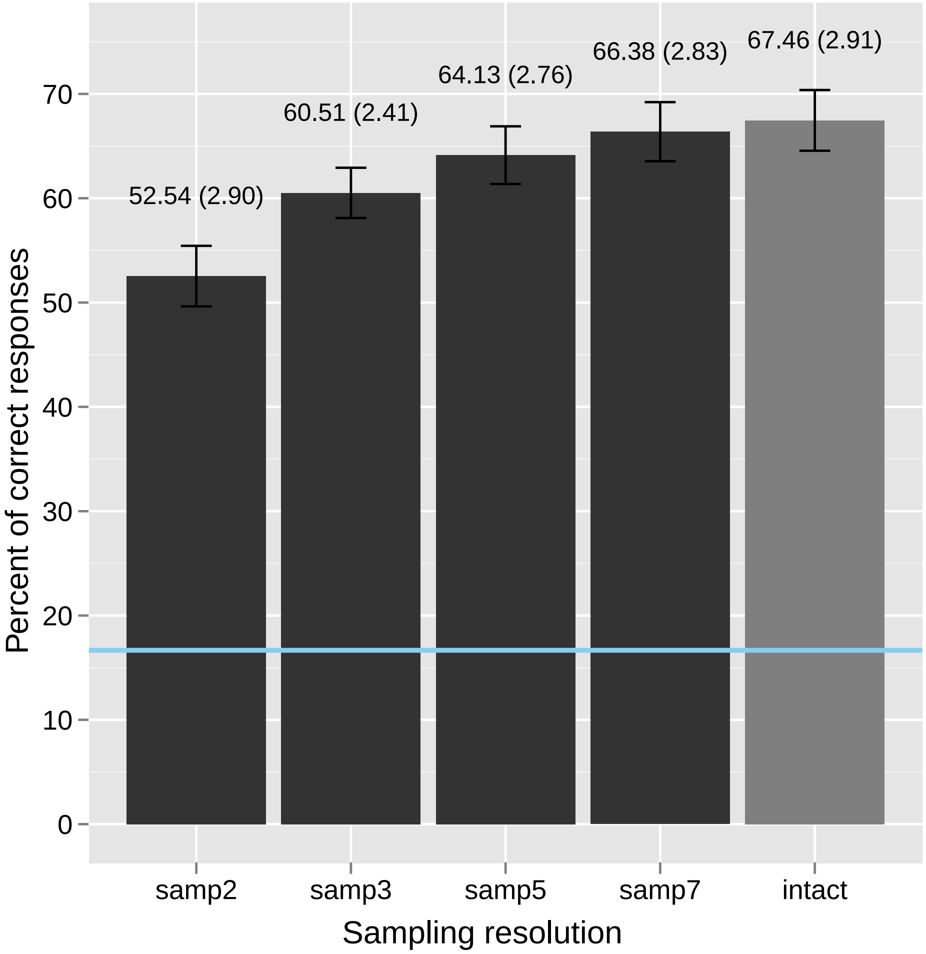


Figure 4.3: Comparison of Cantonese native listeners' overall tonal identification accuracy for different sampling resolutions. Error bars show ±1SE. Tonal identification accuracy was maintained from the intact signal down to 5 samples/syllable. For all sampling resolutions, performance was also well-above chance (the horizontal line shows identification accuracy for at-chance performance (1/6, 17%)).

4.3.2.2 Identification accuracy of individual tones conditioned on sampling resolution

Figure 4.4 displays tonal identification accuracy for each individual tone conditioned on SAMPLING RESOLUTION; for the tones in the higher pitch range of the Cantonese tonal inventory—Tones 55, 25, and 33—as well as Tone 21, tonal identification accuracy was generally high and little affected by sampling resolution down to 3 samples/syllable, but for two tones in the low pitch range, Tones 23 and 22, accuracy was generally low, and accuracy for Tone 23 was also particularly sensitive to decreasing sampling resolution. The bold-faced columns in Table 4.2, the confusion matrix aggregated over listeners conditioned on TONE and SAMPLING RESOLUTION, show that Tones 55, 25, 33, and 21 were identified with between 70 and 90% accuracy in the intact condition and around 60 to 90% accuracy and around 50 to 70% accuracy in the 3- and 2-sample conditions, respectively. In contrast, accuracy for Tones 23 and 22 was only 45 to 50% in the intact condition, dropping to 30 to 40% and to just below 30% for the 3- and 2-sample conditions, respectively.

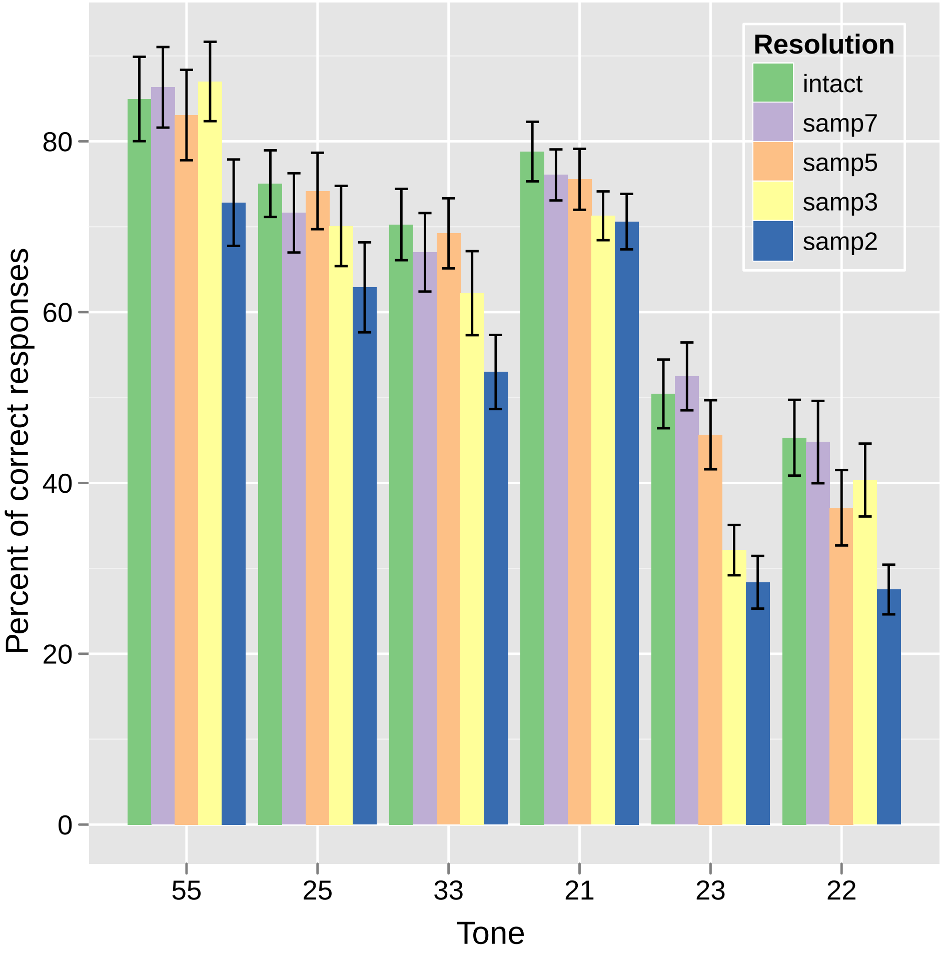


Figure 4.4: Cantonese native listener's tonal identification accuracy for each of the six tones conditioned on SAMPLING RESOLUTION. Accuracy for all tones except Tone 23 was limited in sensitivity to SAMPLING RESOLUTION down to the 3-sample condition. Tones 23 and 22 were identified with strikingly lower accuracy overall than the other tones.

Actual	**Response**					
	55	25	33	21	23	22
55						
samp2	**72.82**	4.27	9.23	1.37	9.74	2.56
samp3	**87.01**	0.85	5.30	0.34	4.96	1.54
samp5	**83.08**	1.37	6.84	0.17	6.32	2.22
samp7	**86.32**	0.51	4.96	0.68	4.62	2.91
intact	**84.96**	1.37	6.32	0.68	5.81	0.85
25						
samp2	2.05	**62.91**	9.23	3.59	18.29	3.93
samp3	1.03	**70.09**	5.30	1.54	20.51	1.54
samp5	0.85	**74.19**	4.10	1.20	18.29	1.37
samp7	0.85	**71.62**	6.32	0.17	18.97	2.05
intact	0.34	**75.04**	3.25	1.03	18.97	1.37
33						
samp2	5.98	5.98	**52.99**	3.25	17.95	13.85
samp3	9.40	2.56	**62.22**	2.91	9.06	13.85
samp5	8.89	2.91	**69.23**	2.74	6.50	9.74
samp7	8.72	2.39	**67.01**	2.22	8.03	11.62
intact	9.06	1.20	**70.26**	1.71	4.96	12.82
21						
samp2	1.88	9.06	4.96	**70.60**	9.40	4.10
samp3	1.03	9.40	4.10	**71.28**	8.03	6.15
samp5	0.85	8.89	3.08	**75.56**	7.01	4.62
samp7	1.03	7.69	3.25	**76.07**	6.50	5.47

intact	1.03	5.30	4.62	**78.80**	4.96	5.30
23						
samp2	2.91	23.93	17.61	13.85	**28.38**	13.33
samp3	3.08	17.61	17.26	12.82	**32.14**	17.09
samp5	1.37	17.95	16.41	6.84	**45.64**	11.79
samp7	2.56	20.34	12.65	4.27	**52.48**	7.69
intact	2.91	17.44	14.53	4.27	**50.43**	10.43
22						
samp2	3.08	9.06	17.78	19.66	22.91	**27.52**
samp3	3.93	4.27	20.51	18.80	12.14	**40.34**
samp5	4.10	5.64	17.78	17.78	17.61	**37.09**
samp7	3.08	3.42	19.15	13.16	16.41	**44.79**
intact	4.27	4.10	18.97	11.28	16.07	**45.30**

Table 4.2: Confusion matrices for each tone for the different sampling resolution conditions.

Results from Tukey tests for comparisons between the degraded and intact conditions (Table 4.3) showed that identification accuracy was not significantly different between the 5- and 7-sample conditions and the intact condition for any tone. For the 3-sample condition, only Tones 21 and 23 were identified with significantly lower accuracy than in the intact condition. However, for all tones but Tone 25, identification accuracy was significantly different in the 2-sample condition from that in the intact condition, and for Tone 25, there was a trend ($p = 0.069$) for a difference in accuracy.

	55	25	33	21	23	22
SAMP2	**0.022**	<u>0.069</u>	**0.002**	**0.016**	< **0.0001**	**0.001**
SAMP3	0.54	0.58	0.11	**0.026**	< **0.0001**	0.39

Table 4.3: Summary of comparisons of identification accuracy conditioned on tone in the 2- and 3-sample conditions with the intact condition. There were no significant differences for the comparisons for 5- or 7-sample conditions so they are not displayed. The p-values estimated from Tukey post-hoc comparisons are given; in all cases of a significant difference, accuracy was lower in the degraded condition than in the intact condition.

Results from Tukey tests for comparisons of accuracy between the degraded conditions (Table 4.4) showed no significant differences between the 3-, 5-, and 7-sample conditions except for Tone 23, which showed significant differences in accuracy between all degraded conditions except between the 2- and 3-sample and between the 5- and 7-sample conditions. There were significant differences between accuracy in the 2-sample condition and accuracies for all other conditions for Tones 55 and 33 and between the 2-sample and 3- and 7-sample conditions for Tone 22. For Tones 25 and 21, there were no significant differences in accuracy

Resolutions	55	25	33	21	23	22
SAMP3 - 2	**0.010**	0.49	**0.031**	0.99	0.71	**0.012**
SAMP5 - 2	**0.018**	0.097	**0.004**	0.16	**0.001**	0.10
SAMP7 - 2	**0.015**	0.19	**0.002**	0.11	$<$ **0.0001**	**0.002**
SAMP5 - 3	0.50	0.78	0.29	0.26	**0.002**	0.92
SAMP7 - 3	0.98	0.99	0.40	0.18	$<$ **0.0001**	0.65

Table 4.4: Summary of comparisons for identification accuracy conditioned on tone between degraded SAMPLING RESOLUTION conditions. The p-values estimated by post-hoc Tukey tests is given; in all cases of a significant difference, accuracy was lower in the more degraded condition than in the less degraded condition.

between any degraded conditions.

The confusion matrices for the six tones conditioned on SAMPLING RESOLUTION in Table 4.2 and Figure 4.5 show what confusability patterns caused the significant drops in accuracy as SAMPLING RESOLUTION decreased. The tones identified with lowest accuracy, Tones 23 and 22, were most confusable with Tone 25 and Tone 23, respectively, overall. This pattern of confusability followed a general trend shown in other tones: on the one hand, the two rises Tone 25 and Tone 23 were consistently most confusable with one another (around 20% for every condition), and Tone 33 was most confusable with the other level tones Tones 22 and 55 down to the 5-sample condition, intuitively explainable as confusions between pitch contours of the same direction (e.g. as confusions along the "direction" dimension in Gandour (1981) multidimensional scaling perceptual space for Cantonese tones).

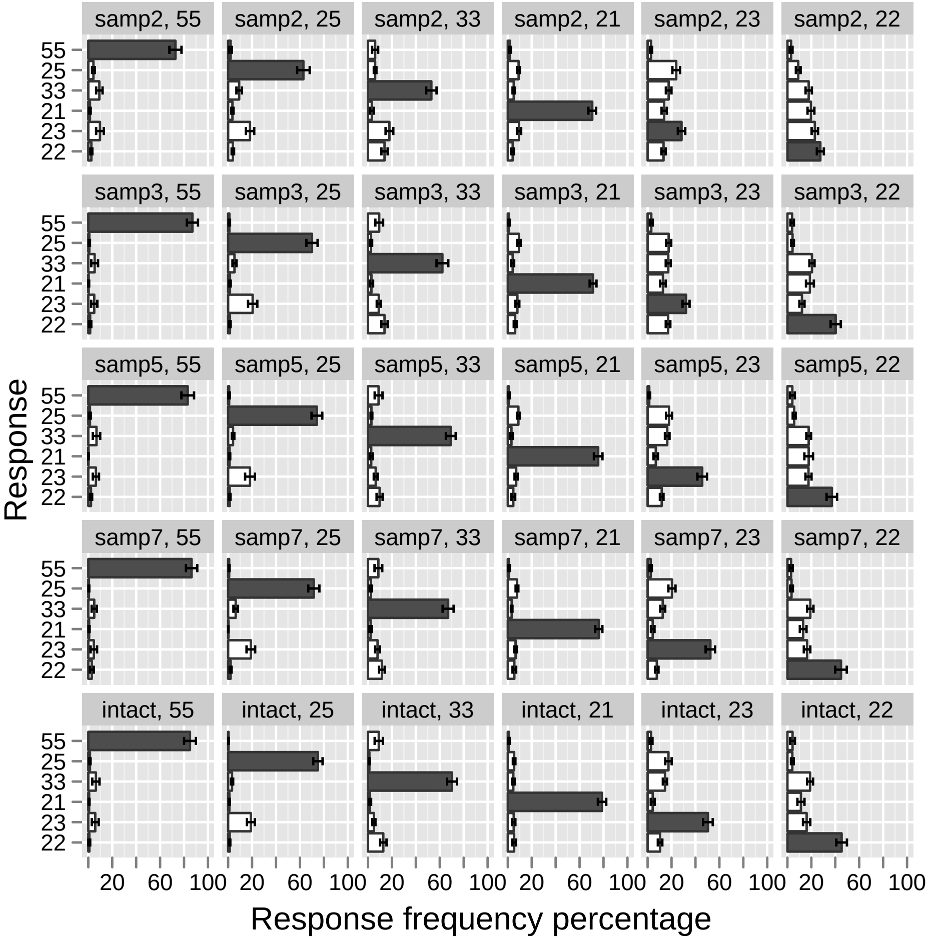


Figure 4.5: Visualization of confusion matrices for each tone for different sample resolutions. The confusion matrices for a given tone run down a single column, with the confusion matrix for each sampling resolution condition in a different row. The horizontal bars display the percentage of responses given for each of the six different tones; the dark grey bars indicate correct responses, and the error bars show ±1SE.

On the other hand, many confusability patterns mixed level and contour tones and rises with falls, especially with decreasing sampling resolution. Tone 23 may have been most confusable with Tone 25, but showed a 6% increase in confusability with the Tone 21 fall between the 5- and 3-sample conditions. Tone 22 was most confusable not only with the level tone Tone 33 as mentioned above, but also the rise and fall Tones 23 and 21, with jumps of 7-8% in confusability with these contour tones between the intact and 2-sample conditions. Tone 33 may have been most confusable with other level tones down to the 5-sample condition, but was most confusable with Tone 23 in the 2-sample condition, 13% and 9% more confusable than in the intact and 3-sample conditions, respectively. Tone 55 was consistently as confusable with Tone 33 as Tone 23.

4.3.3 Discussion

Results from the human perception experiment support the hypothesis that fine-grained temporal resolution of the unfolding speech signal in human perception is not necessary for tonal identification. First, even down to 2 samples per syllable, with less than a quarter of the duration of the speech signal available to the listener, overall tonal identification accuracy was well above chance. Moreover, identification accuracy was maintained from the intact to the 5-sample condition for every tone, and accuracy between the 3-sample and intact conditions differed significantly only for 2 of the 6 tones. The programming error in the stimuli noted earlier that introduced longer noise intervals at stimulus offset in the 2- and 3-sample conditions (27 and 7 ms, respectively) does not weaken the result that fine-grained resolution is not necessary for good separability of Cantonese tones in perception by native listeners. The effect of the error could have only been in the direction of decreasing accuracy at the two lowest SAMPLING RESOLUTION conditions due to interference and/or memory effects.

Our results showing significant decreases in tonal identification accuracy from the intact condition to the 3-sample condition for two of the tones and to the 2-sample condition for almost all tones should not be literally interpreted as providing evidence that the perceptual space for Cantonese tones has around 3 to 5 samples per syllable. First, the error that added more noise at the offset for the 2- and 3-sample conditions which may have contributed to the decrease in identification accuracy for those conditions. In addition, the significantly lower accuracy (and trend for lower accuracy for Tone 25) in the 2-sample condition compared to that in the intact condition may have been due in part to a lack of perceptual continuity caused by the long duration of the interrupting noise intervals for our particular experiment design. (Recall that manipulation of the sampling resolution involved an increase of the duration of the noise interval as sampling resolution decreased.) In support of this conjecture, Dannenbring (1976) showed that in nonspeech, for pure tones of 250 ms in duration interrupted by white noise, the mean continuity threshold between perceived continuity and discontinuity due to the interrupting noise was around 80 and 100ms for steady state tones and tone glides, respectively. This indicates that the 90-ms noise interval duration in the 2-sample condition may have been close to the auditory threshold for perceiving continuity in our stimuli, which were 241 ms in duration.

More generally, the experiment was not designed to test *where* samples could be taken (with alignment specified, for instance, with respect to segments or syllables) to maximize tonal separability in the perceptual space, since the samples were uniformly distributed over the syllable. For our purposes, it was sufficient to show that there exists *some* set of samples that maintains tonal separability of the intact condition even under coarse sampling resolution, even if the set of samples provided in the stimuli for a given sampling resolution was not the set that optimized tonal separability. It is possible that even a single sample per syllable taken at a timepoint where the separability of the tones is optimal over the syllable could be sufficient for discriminating tones in natural conditions, although we did not test a 1-sample condition because perceptual continuity of the stimuli would have been lost under the long noise duration intervals required for such a condition with our experimental design. Khouw and Ciocca (2007) found that f0 change over the 6th and 7th out of 8 subsyllabic segments accounted for about 70% of the variance in a Cantonese tonal identification perception experiment of isolated monosyllables, and in Cantonese (Li et al., 2002, 2004; Wong, 2006) as well as other (South)east asian languages (Mandarin: (Xu, 1997), Thai: (Gandour et al., 1992), Vietnamese (Han and Kim, 1974)), it has been reported that rightward (carryover) coarticulation is stronger than leftward (anticipatory) coarticulation, so that tones in connected speech might be maximally separated near the offset of the syllable.

While understanding variation in the identification accuracy of individual tones overall and as a function of SAMPLING RESOLUTION is not crucial for addressing our research question, it is of interest to for insight into how and why sampling resolution affects tonal identification accuracy. The most striking variation in overall identification accuracy was the low accuracy for Tones 22 and 23: even in the intact condition, overall identification accuracy was 43% for these two tones, compared to 77% for the other four. We conjecture that the low accuracy for Tones 22 and 23 was due in part to tonal mergers in some of the listeners, because these tones were particularly confusable with other tones in the context of the mid level tones flanking the target tone to be identified, and also possibly due to relatively lower lexical frequency of the characters used for these two tones in the identification task.[8] We also conjecture that Tone 21 was identified with high accuracy unlike its close neighbors Tones 22 and 23 not only because it occupied a lower part of the pitch range than them, but also because of creaky voice quality cues, since Yu and Lam (2011) showed that the presence of creaky voice cues can boost Tone 21 identification accuracy in Cantonese tone perception.

It has been reported that it is not unusual for Cantonese native speakers in Hong Kong in their 20s, the population from which our subjects were sampled, to have tonal mergers, especially between Tones 33/22, Tones 25/23, and Tones 21/22, e.g. Mok and Wong (2010a,b) and references therein. We did not screen our listeners for mergers, but post-hoc inspection of individual results suggests that some of the listeners may have had mergers since they showed systematic response biases across resolution conditions. There were around 5 subjects who almost never gave correct responses for Tone 23 regardless of condition, and of these, two gave mostly Tone 25 responses in the intact condition. There were also around 5 subjects who rarely gave correct responses for Tone 22 regardless of condition, and of these, 2 gave mostly Tone 21 responses. Thus, there may have been subjects with Tone 25/23 and Tone 21/22 mergers.

However, the other subjects who performed poorly overall on Tones 22 and 23 gave incorrect responses distributed over a mix of tones rather than mostly over a single tone, and some subjects identified Tone 22 as mostly Tone 23 in the intact condition. We therefore conjecture that the context in which the stimuli

[8] As a rough estimate of lexical frequencies of the six orthographic characters used in the identification task, we used the frequencies of the Mandarin cognates, [wei], in the character frequency list of Modern Chinese from Da (2004). Counts from that text corpus indicated the following relative frequency percentiles, from the most to least frequent character used to represent the tones: T2 (26), T4 (21), T1 (20), T5 (9), T6 (3), T3 (3). However, T3 identification accuracy was similar to that of the T2, which had the highest character relative frequency percentile, and more than 20% higher than that of T5 and T6, while having a character relative frequency percentile as low as T6.

were presented—with flanking mid tones (Tone 33)—may have been a context in which Tones 22 and 23 were particularly confusable with other tones—more so than the other four tones—since they were mid-low tones that didn't deviate much from Tone 33. In support of this, even the 5 listeners who identified tones in the intact condition with around 90% accuracy, 4 of whom identified tones in the 2-sample condition with 65-73% accuracy, showed markedly lower accuracies for Tones 22 and 23 relative to the other tones in the 2-sample condition. They also exhibited confusion patterns in the 2-sample condition similar to those in the intact condition for listeners who performed poorly on Tones 22 and 23 overall: confusion of Tone 23 with Tone 25 and Tone 22 with Tone 23.

These and the other patterns of confusion in the perception experiment showed that while there were overall, consistent confusions between rises and between level tones as has been shown in experiments with isolated monosyllables (Fok, 1974; Gandour, 1981; Khouw and Ciocca, 2007), there was also much confusion between tones with different contour shapes and directions in our connected speech stimuli, especially as sampling resolution decreased. These confusions may have arisen from uncertainty about the magnitude of rises and falls in pitch and their alignment with segmental material in the face of degraded sampling resolution, particularly in the Tone 33 - Tone X - Tone 33 context used in the experiment.

We could check the role of acoustic separability in the particular context used in our experiment indirectly by analyzing confusion matrices from previous perceptual experiments in the literature, but the tonal contexts from those are generally an isolation/citation context, e.g. Fok (1974); Khouw and Ciocca (2007). One of the closest matches for context comes from a condition in Ma et al. (2005, Table 4), a Tone 33 - Tone X - Tone 33 context, and the confusion matrix shows relatively low accuracy, close to 50% for Tone 25 and Tone 22, with all other tonal accuracies around 90% or above. These results match the low Tone 22 accuracy in our experiment, but not other results. We directly tease apart the role that acoustic separability alone has to play in our experimental stimuli using computational modeling in the following section.

4.4 Computational modeling

The perception experiment (§4.3) showed that native listeners can classify Cantonese tones at well above chance levels—some listeners with accuracy around 70%— even with as few as two 30ms samples per syllable, with less than a quarter of the duration of speech signal available. However, it did not tell us *how* listeners might be doing this. In each sample, there are an infinite number of acoustic parameters available to the listener. Moreover, there is an unbounded range of evidence outside the speech signal that the listener could bring to bear on the classification task, including the lexical biases discussed in §4.3.3.

To gain insight into what the listeners could be attending to, we used computational methods to model the classification problem, defined under precise assumptions. Our purpose was to determine: *given a minimal acoustic parameterization of the speech signal and abstracting away from other sources of evidence, could the stimuli in our experiment be classified as accurately under the degraded* SAMPLING RESOLUTION *conditions as with the intact speech signal?* To this end, we defined the raw acoustic parametrization of the stimuli to be: (i) the set of mean f0 values from each sample, for the degraded conditions, or (ii) f0 tracks extracted with a 10ms frameshift, for the intact condition. We chose linear support vector machines (SVMs) as our classifiers (Vapnik, 1995; Cortes and Vapnik, 1995; Burges, 1998). SVMs are well-understood and widely used in machine learning and have been used in automatic tonal recognition, e.g. Levow (2005); Peng and Wang (2005).

We sketch a geometrical characterization of how they work for the binary case, e.g. for two tone classes, following Bennett and Bredensteiner (2000). Call the two classes Class A and B. Each stimulus is parameterized as a real-valued p-dimensional vector and labeled as belonging to either Class A or B. Thus, the Class A and B stimuli sets each comprise a set of points in $\mathbb{R}^p$. The SVM algorithm is a way to determine an optimal decision rule to assign a class label to a stimulus. A linear SVM determines a $p-1$ dimensional separating hyperplane as a decision boundary in the parameter space, i.e. a 1-dimensional line for stimuli parameterized in 2-D space, $\mathbb{R}^2$. The SVM algorithm chooses the optimal separating hyperplane to be the one that maximizes the distance from the hyperplane to the Class A and Class B sets.

Which hyperplane is this? Take the convex hulls of the Class A and Class B sets, the set of points enclosed in the tightest rubber band one can stretch around the Class A and B sets, respectively. The optimal hyperplane bisects and is orthogonal to the line segment between the two closest points of the convex hulls (Boyd and Vandenberghe, 2004, p. 46-49). If Class A and B are linearly inseparable, i.e. if their convex hulls overlap, then a soft margin SVM algorithm can be used, which allows for some points to be on the wrong side of the margin in determining the optimal separating hyperplane, and a soft margin parameter is tuned to balance the tradeoff between maximizing the margin and minimizing classification error.

We desire the determined classification rule to generalize beyond the training data used to choose it. Thus, evaluation of classifier performance is done by determining classification accuracy on test data, data not in the training data set: in this study, we trained classifiers on stimuli from a subset of 4 of 5 of the speakers and tested the classifiers on the withheld speaker.

4.4.1 Methods

We implemented the linear SVMs using LIBSVM (Chang and Lin, 2001). Because the SVM algorithm involves calculating Euclidean distances in the parameter space, it is necessary to scale the data, so that parameters with a greater range do not dominate the direction of the optimal separating hyperplane relative to parameters with a smaller range, and it also necessary for the training and test data to be scaled in the same way. Thus we chose to parameterize the stimuli using z-score standardized log-transformed f0 rather than f0 in Hz (§4.2.3), cf. (Levow, 2006, §2.3).

We used the default treatment of multiclass classification in LIBSVM, which decomposes the 6-way Cantonese tone classification problem as $\binom{6}{2} = 15$ binary classification sub-problems, and then uses a voting strategy to combine the 15 decisions. For each sampling resolution condition, we used 5-fold cross-validation and partitioned our data into 5 folds, one fold per speaker. Rotating across the folds, a single fold (18 tritones, 1 speaker) was withheld as test data, and the remaining four folds ($4 \times 18 = 72$ tritones, 4 speakers) were used as training data. The soft margin parameter was chosen for each rotation using 5-fold cross-validation on the training data. All classification results, unless otherwise indicated, are averaged across the results from the 5 rotations, and standard error for classification accuracy is calculated from the variance of the accuracy over the 5 folds.

4.4.2 Results

Overall, classification accuracy was higher in the SVM classification than in human listeners, and there was no significant difference between the SVM classification accuracy for any of the sampling resolution conditions (Table 4.5), based on Bonferroni-corrected t-tests paired by fold. Thus, SVM classification accuracy with as few as 2 real values per syllable (standardized log-transformed f0), 6 in total over the tritone, was not statistically different from accuracy with real values sampled every 10ms, 69 in total over the tritone: an order of magnitude in the number of real-valued parameters had no effect on classification accuracy! Moreover, classification accuracy was well above chance for all sampling resolution conditions, as in the human perception experiment.

Sampling resolution	Percent correct (SE)
samp2	76.67 (7.93)
samp3	83.33 (5.56)
samp5	77.78 (6.09)
samp7	76.67 (6.43)
10ms frameshift	81.11 (5.98)

Table 4.5: SVM tonal classification accuracy conditioned on sampling resolution, aggregated across the speaker folds. SE, in parentheses, is derived from between-fold variance.

Confusion matrices for the 10ms frameshift and 2-sample/syllable conditions are given in Tables 4.6 and 4.7. There were no classification errors for Tone 55 or 33, for any sampling resolution. The Tone 25 rise was classified with the highest accuracy after that, around 90%, and was confused with Tone 23, the other rise. Tone 22 was classified overall with around 70-80% accuracy and was highly confusable with Tone 23 across sampling resolution conditions. Tone 21 and Tone 23 were classified with the lowest accuracy overall, around 40-70%, and were both confused with a mix of tones. Tone 21 was mostly confused with Tone 23 with a 20% increase in confusability with Tone 22 between the 10 ms frameshift and 2-sample conditions, and Tone 23 was mostly confused with Tone 22 and Tone 21.

These results are similar in some ways to the perception experiment results. For both human listeners and machine, Tones 55, 25, and 33 were classified with high accuracy, and Tone 23 with low accuracy. However, for humans, Tone 21 was classified with high accuracy and Tone 22 with low accuracy, but for the SVMs, Tone 21 accuracy was relatively low, and Tone 22 accuracy relatively high. In addition, unlike for human listeners, there was no pattern of sharp drops in classification accuracy between the 3- and 2-sample conditions except for Tone 21, which had a 13% increase in Tone 22 confusability.

Actual	**Response**					
	55	25	33	21	23	22
55	**100.00**	0.00	0.00	0.00	0.00	0.00
25	0.00	**86.67**	0.00	0.00	13.33	0.00
33	0.00	0.00	**100.00**	0.00	0.00	0.00
21	6.67	0.00	6.67	**66.67**	13.33	6.67
23	0.00	0.00	6.67	20.00	**53.33**	20.00
22	0.00	0.00	0.00	0.00	20.00	**80.00**

Table 4.6: Confusion matrix from SVM classification for the 10 ms frame shift condition.

Actual	**Response**					
	55	25	33	21	23	22
55	**100.00**	0.00	0.00	0.00	0.00	0.00
25	0.00	**93.33**	0.00	0.00	6.67	0.00
33	0.00	0.00	**100.00**	0.00	0.00	0.00
21	0.00	6.67	6.67	**40.00**	20.00	26.27
23	0.00	6.67	6.67	13.33	**53.33**	20.00
22	0.00	0.00	0.00	0.00	26.67	**73.33**

Table 4.7: Confusion matrix from SVM classification for the 2-sample condition.

4.4.3 Discussion

Our computational modeling showed that given only a minimal acoustic parameterization of the speech signal, with one real-valued parameter—standardized log-transformed mean f0—per sample, the stimuli in the perception experiment could be classified as accurately under every degraded sampling resolution condition, with a lower bound of 6 real values for parameterization (2 samples/syllable), as with 69 real values. The tonal identification accuracy from computational modeling, in the high 70s to 80%, was close to that from the supratone Cantonese automatic tonal recognizer (Qian et al., 2007), which uses Gaussian mixture models based tritone models over a parameter set of 3 averaged f0 values per syllable and had an accuracy of 75.59% on a much larger and variable corpus than ours. Thus, acoustic information in the signal alone is sufficient for explaining how accurate tonal classification with less than a quarter of the speech signal duration available *could* be maintained. This is not to say that listeners were computing standardized log-transformed mean f0s over each sample and making tonal decisions based on SVM algorithmic procedures! We merely suggest that computational modeling shows that the information needed for accurate tonal classification in the face of sparsely sampled data is present in the acoustic signal. We discuss the interplay of assumptions in computational modeling and what humans actually do further in §4.5.

In fact, it is clear that our results from computational modeling differ significantly from the human perception results in a few ways. First, human performance dropped as sampling resolution did, but machine performance did not. We conjecture that part of the reason for this difference is that humans were coming up against the limit of perceptual continuity as the noise interval durations increased, as discussed in §4.3.3, while the effect of perceptual continuity was not computationally modeled. Additionally, we did not model tonal mergers in the machine classification, and some human listeners may have had tonal mergers. The z-score preprocessing of the data, as a speaker normalization procedure, may have also been the reason that the level tones Tone 55 and Tone 33 were identified with perfect accuracy and Tone 22 with high accuracy for all resolution conditions by machine, but not by humans.

Second, identification accuracy for Tone 21 was relatively high for humans, but low for machine, and accuracy for Tone 22 was relatively low for humans, but high for machine. These discrepancies are informative for insight into the human perception experiment. Most of the unvoiced frames in the RAPT f0 extraction came in Tone 21 stimuli, since Tone 21 realization frequently had nonmodal phonation, low amplitude, and even intervals of silence. The parameterization for computational modeling did not capture this, and while there are more sophisticated rules for estimating the pitch percept in the presence of nonmodal phonation for the human listeners than the simple linear interpolation used over voiceless frames that we used, cf. the aberrant pitch contours for Tone 21 in Figure 4.2, the poor accuracy for Tone 21 identification by machine nevertheless suggests that a parameterization of the speech signal that references voice quality, beyond f0, is needed for both higher classification accuracy of T4 by machine, and for modeling what humans are doing. In support of this, as discussed in §4.3.3, Yu and Lam (2011) showed that Cantonese listeners use voice quality parameters in tonal perception, and that creaky voice cues improve their Tone 21 identification accuracy.

The discrepancy between human and machine for Tone 22 accuracy may have been due to lexical biases that we abstracted away from in the computational modeling. The estimated low relative frequency of the Tone 22 character (cf. §4.3.3) may have biased listeners against Tone 22 responses, as the computational modeling suggests that the acoustic information available could have resulted in better performance relative to T5 identification.

4.5 General discussion

In summary, we have shown that fine-grained temporal resolution in the speech signal is not necessary for accurate tonal perception within the contexts of our human tonal perception experiment and computational modeling. The task of the listeners in the perception experiment was to identify the middle tone in nonce Cantonese tritones from five speakers, where the first and third members of the tritones were fixed as mid level Tone 3(3), and where the sampling resolution in the speech signal was systematically degraded by interrupting the signal with uniformly distributed noise intervals. In computational modeling of tonal classification of the experimental stimuli using linear support vector machines, the speech signal was parameterized as z-score log-transformed f0 values standardized over speaker-specific means and standard deviations, extracted at a 10 ms frame shift, and averaged over the same frames that were available to the listeners.

Humans as well as linear support vector machines classified the tones at well above chance levels even at a temporal resolution of 2 samples/syllable, a condition where less than a quarter of the duration of the speech signal was accessible. For human listeners, identification accuracy at 5 samples/syllable was not statistically distinct from accuracy with the intact speech signal, and accuracy at 3 samples/syllable was also not statistically distinct from the intact condition for all but 2 of the 6 tones. The noise interval durations in the 2-sample condition may have been close to the upper limits of perceptual continuity, which may have been part of the reason that five tones were identified with significantly less accuracy and one tone was identified with a trend for less accuracy in that condition than in the intact condition. For machine classification, accuracy was not statistically different between using a parameter set of 69 standardized f0-based values and the parameter sets for any degraded conditions down to the 2-sample condition analogue, with a parameter set of 6 f0-based values over the tritone, showing that even with sparse temporal resolution, acoustic information alone is sufficient for accurate tone classification.

As discussed in §4.1, this negative result that fine-grained attention is not necessary for good separability of Cantonese tones in human listeners and by machine classification is consistent with rigidity in the structure of the class of maps from the speech signal to lexical tone categories in human languages—it is consistent with finite VC dimension in tonal maps in human languages. Additionally, our perceptual results also bear on tonal recognition, especially since we set up the manipulation of sampling resolution for the listeners in a way consistent with acoustic feature extraction in tonal recognition, with a uniform sampling rate over multiple syllables (tritones) to provide contextual information, with syllables duration fixed to a constant to simulate time normalization in feature extraction. Our perceptual results showing the sufficiency of coarse-grained sampling resolution for human listeners complement Tian et al. (2004)'s automatic tonal recognition study of Mandarin that found that sparse temporal resolution, with 4 samples/tone, can outperform fine-grained sampling with 1 sample/10 ms and help motivate the coarse-grained frameshift in automatic tonal recognizers that has been used in many recent studies such as Peng and Wang (2005); Qian et al. (2007); Wang and Levow (2008); Zhou et al. (2008).

Since automatic tonal recognition and more generally, automatic speech recognition has not reached levels of human performance, one apparent remaining puzzle arising from our proof-of-concept machine classification experiment is the gap between human and machine performance: while the best human performance was 67.71% accurate, with no noise in the signal, SVM performance was around 80%. We already discussed in §4.3.3 that some of this gap could be attributable to tonal mergers in listeners. Moreover, it is important to remember that the human and machine tasks were incomparable, and so the raw accuracy values obtained in the two tasks are also incomparable. The machines were trained to just stimuli from the experiment—tritones, with the first and third tones fixed to mid-level tones—and then tested on stimuli from the experiment. The humans had a lifetime of training on a huge variety of contexts and then were tested on stimuli in a simulated "pure speech" context, with very limited top-down information available—a test context that was surely never encountered in their lifetime training.

With that comparison of training in mind, one can interpret the high performance of the machines as due in part to overfitting: should the trained classifiers be asked to identify tones in contexts other than that of the stimuli in this experiment, they would likely do poorly. In addition, the machines received the stimuli in a representation that may have been unlike how humans perceive the stimuli. The preprocessing of the stimuli pitch contours as log-transformed f0, standardized by speaker pitch range, gave the machines different information than was available to the speakers, which appears to have been particularly informative for level tone identification.

While neither the perceptual experiment nor the computational modeling in this chapter tells us what restrictive structure might be present in the class of tonal maps in human languages, our results highlight a puzzle: this study establishes that fine-grained temporal resolution in the speech signal is not *necessary* for distinguishing tonal concepts in natural languages, yet humans are sensitive to fine-grained temporal resolution. Krishnan et al. (2005) even show that the frequency following response in the brainstem, thought to encode pitch in humans, shows higher-fidelity pitch tracking in Mandarin speakers than English speakers and suggest that native speakers of tone languages may be tuned by their language input to be more sensitive to fine-grained temporal resolution in pitch encoding. With such fine-grained resolution, the potential number of distinctions that could be drawn as tonal concepts, over the time series of even just a single parameter, could explode.

Yet, tonal maps of the world's languages seem to have a tendency to be simple in that the tonal concepts of a particular language can be well-separated in purely acoustic spaces of low dimensions, as evidenced by the results of this paper and other studies parameterizing the speech signal as an f0-based time series sampled with sparse temporal resolution, cf. §4.1, or other low-dimensional representations of f0, e.g. piecewise-linear approximations (refs. in Hirst and Espesser (1993, 75), Li and Lee 2007) quadratic splines (Hirst and Espesser, 1993), or the orthogonal Legendre polynomial basis set (Kochanski et al., 2005). At this point in time, it is not clear whether the simplicity of tonal systems in natural languages is due to: (i) a natural class of learnable tonal systems that may, as a class, have finite VC dimension (Vapnik, 1995) or (ii) a strong probabilistic tendency for tonal systems in natural language to be simple, despite being drawn from a class of possible tonal systems in natural languages that may be unbounded in complexity, for instance, in having infinite VC dimension.

Discovering what restrictive structure might be present in tonal maps in human languages will take both behavioral and neurological experiments with humans and computational modeling to tease apart what humans are doing in those experiments. The fit between human experiments and computational modeling in this study is crude at best, and it is our hope that future work can simultaneously ground modeling assumptions more carefully based on what we learn about human cognition, and sharpen the questions we ask and conclusions we can draw from human experiments with formal computational perspectives.

CHAPTER 5

Cross-linguistic data for studying tonal representations from Bole, Mandarin, Cantonese, and Hmong

5.1 Introduction

In this chapter, we describe exploratory data analysis comparing the separability of tonal concepts in different parameter spaces in single-speaker spaces, using cross-linguistic data from our sample of tone languages: Bole, Mandarin, Cantonese, and Hmong. Unlike in the preceding three chapters, our analyses are entirely based on computational modeling. We define separability using both classification accuracy calculated from machine classification and geometric visualization of the overlap of the convex hulls of tonal concepts, as introduced in the discussion of the support vector machine algorithm in Chapters 2 and 4.

To determine how tones are defined for humans—how they are parameterized—we assume here that a good candidate for a relevant parameter space for humans is one in which separability is maximized, and in particular, one in which tones are (close to) separable, but with a penalty for complexity—here, defined as the number of parameters in the parameter space. For instance, if separability is approximately equivalent in two spaces, but one uses more parameters than the other, then we assume the space with fewer dimensions is the one relevant for tonal representation for humans.

The drawback of lacking human behavioral data here is that the strategy of maximizing separability of tonal concepts is not guaranteed to produce tonal spaces that are close to human tonal spaces. If a principle of maximizing separability, or more generally, a principle of dispersion, were the only principle in play in human tonal spaces, then we would not expect any tonal mergers, for instance, and yet they occur, e.g. in present-day Cantonese (Mok and Wong, 2010a,b). However, this is a reasonable strategy in the face of missing relevant human perceptual data, and the standard one.

The availability of computational modeling allows us to ask some questions that are very difficult, if not impossible, to ask with human behavioral experiments. We sought to generalize both our own results about voice quality and temporal resolution in Cantonese tones and Gauthier et al. (2007)'s claim that f0 velocity alone outperforms f0 values in classifying Mandarin tones (in multi-speaker spaces, not single speaker spaces) to other languages. For a single-speaker tonal space in Bole, Cantonese, Hmong, and Mandarin, we asked:

1. Does phonation interact with f0 parameters in a systematic way? Can we really abstract away from these interactions, if they exist?
2. How is tonal separability affected by the temporal resolution of f0-based parameters—log-transformed f0 values (static) and f0 change (dynamic)?
3. In a d-dimensional parameter space for fixed d, do f0 change parameters yield higher tonal separability than static f0 parameters? How about static and dynamic parameters in combination?

Teasing apart the effect of static vs. dynamic cues is very difficult to approach with human behavioral studies, since f0 values and the change between them are linearly dependent and thus by definition difficult to segregate as cues in a controlled experiment. But this is quite feasible using computational modeling. While Gauthier et al. (2007) used self-organized maps, a kind of neural network approach, to model the unsupervised learning of tone categories and support his claim about f0 velocity in Mandarin, here we use linear classifiers to model supervised learning, as discussed in Chapter 1, since our focus is on characterizing tonal representations. Like Gauthier et al. (2007), though, our stimuli sets are drawn from syllables extracted from connected speech elicited to vary over all licit bitone combinations, and the parameters are drawn only from these syllables and not also neighboring ones.

While we already know that there are Hmong, Mandarin and Cantonese tones that are frequently creaky, as discussed in Chapter 1 and 3, we were interested in following up on our claim from 3 that it is difficult to treat f0 as a separate non-interacting component from other aspects of voice quality: here in this exploratory analysis, we wanted to abstract away from voice quality to work in f0 value-based spaces, but we suspected that it would be difficult to do so.

Further, drawing on our results in Cantonese from Chapter 4, we hypothesized that temporal resolution would not have a large effect on tonal separability within a set of parameters: separability using f0 values only or f0 change values only would not be dramatically increased with higher temporal reslolution. Because our sample of languages included ones with level tone contrasts (Bole, Cantonese, Hmong), we hypothesized that given a fixed number of total parameters for defining tones in these languages, f0 change parameters would not yield higher separability than static f0 parameters. It is important to note that level tones may be realized as contours, as exemplified in the pitch track for a Bole sentence in Figure 5.1, so it is not obvious that f0 change parameters won't separate level tones well.

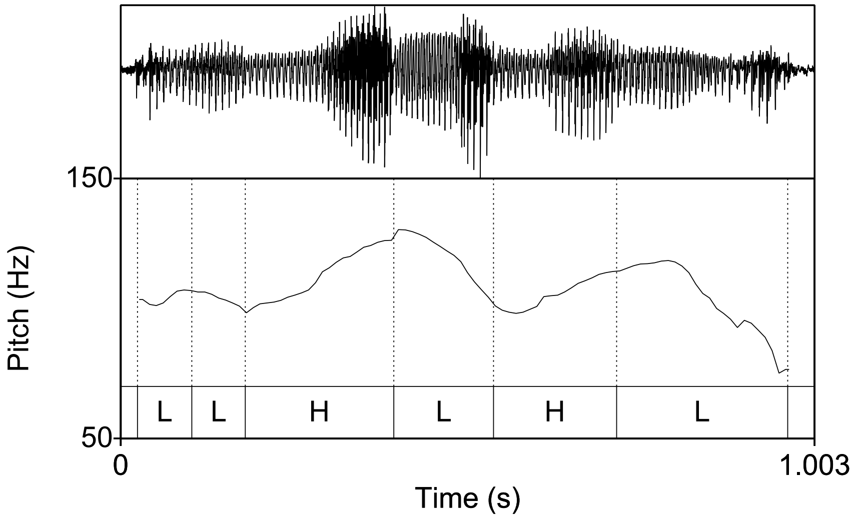


Figure 5.1: A sequence of tones in Bole, a tone language with H and L tones. Sequences of level tones in a level tone language are not necessarily sequences of step functions. Rather, they can show rises and falls due to tonal coarticulation. The sentence is *ànìn némà méngò*, 'The owners of prosperity came back.'

The rest of this chapter first describes the materials and methods used in computational modeling (§5.2) and then includes a note about the interaction of phonation with f0 parameters (§5.3), the results for the effect of temporal resolution on separability (§5.4), the comparison of static and dynamic f0 parameters (§5.5); we conclude with a general discussion (§5.7).

5.2 Materials and methods

5.2.1 Data

For exploratory data analysis, we chose a single male speaker from each language—Bole, Mandarin, Cantonese, and Hmong—from our production data corpora described in Chapter 1. All the speakers had linguistics training and produced tones more accurately and distinctly than other speakers in the corpora, and also were among the least creaky. We chose to model single speaker spaces, since it has been demonstrated that tones in single speaker spaces (speaker-dependent recognition) are much more separable than tones in multiple speaker spaces (speaker-independent recognition): Wong and Diehl (2003) found that in identifying the three level tones of Cantonese in isolation from 7 speakers, listeners were 80.3% accurate when stimuli were blocked by speaker but only 48.6% accurate when stimuli from different speakers were mixed together. With single speaker, relatively dispersed tonal spaces to model, we felt that the assumption of separability of tonal categories would be more well-founded, and with speakers who had little creak in their productions, we thought it would be more reasonable to abstract away from voice quality parameters.

Exemplars of each tone were drawn from productions of sentence-medial bitones over all licit bitone combinations in each language, as described in Chapter 1. The syllable boundaries of each of the bitone members were annotated using Praat (Boersma and Weenink, 2010). Parameters were extracted from monosyllables; no contextual information as in Chapter 2 was included, as was done in Gauthier et al. (2007), to facilitate comparisons with their results.

5.2.2 Parameter sets

We extracted f0 values with RAPT (Talkin, 1995), under default settings, and we created different temporal resolution conditions by averaging and log transforming f0 values over 1, 2, 3, 4, 5, 7, 10, and 12 intervals uniformly partitioning the syllable duration. As in most automatic tonal recognition studies, one effect of this averaging was to throw away information about missing values, say, due to nonmodal phonation. The condition with 1 interval uniformly partitioning the syllable duration, under the averaging performed here, is equivalent to taking the mean over the syllable. Because this is very smoothed compared to extracting a single sample over a smaller temporal window than the syllable, we also compare extracting a single sample from the 5 interval condition to extracting the syllable mean in §5.6. We were primarily interested in comparing 1, 2, 3, 5, and 10 samples, but also calculated f0 means for 4, 7, and 12 samples for f0 change calculations.

We calculated f0 change values from the f0 values as in Gauthier et al. (2007), by taking half the difference between averaged f0 values one interval apart as a kind of smoothing procedure, e.g. for calculating f0 change values for a temporal resolution of 3 parameters per syllable, we took half the difference for the 1st and 3rd, 2nd and 4th, and 3rd and 5th mean f0 values extracted over 5 uniform intervals. By calculating f0 change values for a given temporal resolution from f0 values for a different temporal resolution, we also kept the f0 and f0 change parameters linearly independent when they were combined in the same parameter set. This was necessary for the algorithm used (§5.2.3) to determine a solution for classification, since it relied on calculating matrix inverses. We calculated f0 change parameters, in Hz, for 1, 2, 3, 5, and 10 parameters over the syllable.

The purpose of using log-transformed f0 values but f0 change values in Hz space was to put the two types of parameters on a similar scale; this was irrelevant for our study (cf. §5.2.3), but crucial in Gauthier et al. (2007) for the same reason we standardized parameters for SVM classification in Chapters 4 and 2.

5.2.3 Analysis

We used a different classification algorithm than that for support vector machines for this study because of our interest in dimensionality reduction for the visualization of separability in our exploratory data analysis. The classification algorithm was linear discriminant analysis (LDA) (Duda et al., 2001; Hastie et al., 2009) implemented using the MASS package in R (Venables and Ripley, 2002), which produces classifiers similar to the support vector machines used in earlier chapters: both are linear classifiers; the only difference is in the rule for choosing the optimal separating hyperplane (linear discriminant).

For linear discriminant analysis, a linear combination of the parameters is chosen that maximizes the separability between classes, where separability maximizes the ratio of between-class separability (roughly, the Euclidean distance between class means), and the within-class separability (roughly, the variance within a class). For a 2-class problem, only one linear discriminant is chosen; for multiclass problems with k classes, $k-1$ discriminants are chosen, each orthogonal to the others, ranked in order of how much they contribute to separability. Using the two top-ranked discriminants, one can visualize the 2-D space that produces the most separated classes, under the separability metric defined in LDA. The 2-D parameter space, a near-approximation of the 10-D one, is still defined over the original number of parameters, e.g. 10 parameters if 10 f0 values were sampled over the syllable. For Bole, since there were only 2 classes, we plotted the tones in the LDA-optimized space using density plots, which are like smoothed histograms, in 1-D.

As measures of separability, we used both the leave-one-out cross-validated accuracy of the LDA classifiers and the geometric visualization of 2-D convex hull plots in LDA-optimal spaces. The leave-one-out strategy for evaluating classifier performance is similar to the strategy we used in evaluating SVMs, where we trained on all but one speaker, and then tested the classifier on the held out speaker. Here, we hold out exemplars rather than speakers. The convex hull plots show the convex hull of each tone category in the LDA-optimal spaces: the more overlap there is between convex hulls, the less separated the tone categories are in a space.

5.3 Abstracting away from voice quality cross-linguistically

For this exploratory study, we considered only f0-based parameter spaces. We also selected speakers with minimal creak in their productions. In this section, we show f0 contours from 12 samples over the syllable from each speaker to to demonstrate that nevertheless, there are still hints of creak "perturbing" the f0 contours.

In Bole, a language with a H and L level tone contrast, the f0 contours in Figure 5.2 show that the L tone was generally realized without creak that affected the f0 contour with a few exceptions.

In Mandarin (Table 5.1), a language with T3 known to be frequently creaky, the pitch halving in f0 contours in Figure 5.3 show that indeed, a good proportion of the T3 contours are creaky. Something to note here, too, is that while in isolation, T3 contours are creaky in the middle of the syllable, here, in the sentence-medial context, the creaky region can persist to the end of the syllable.

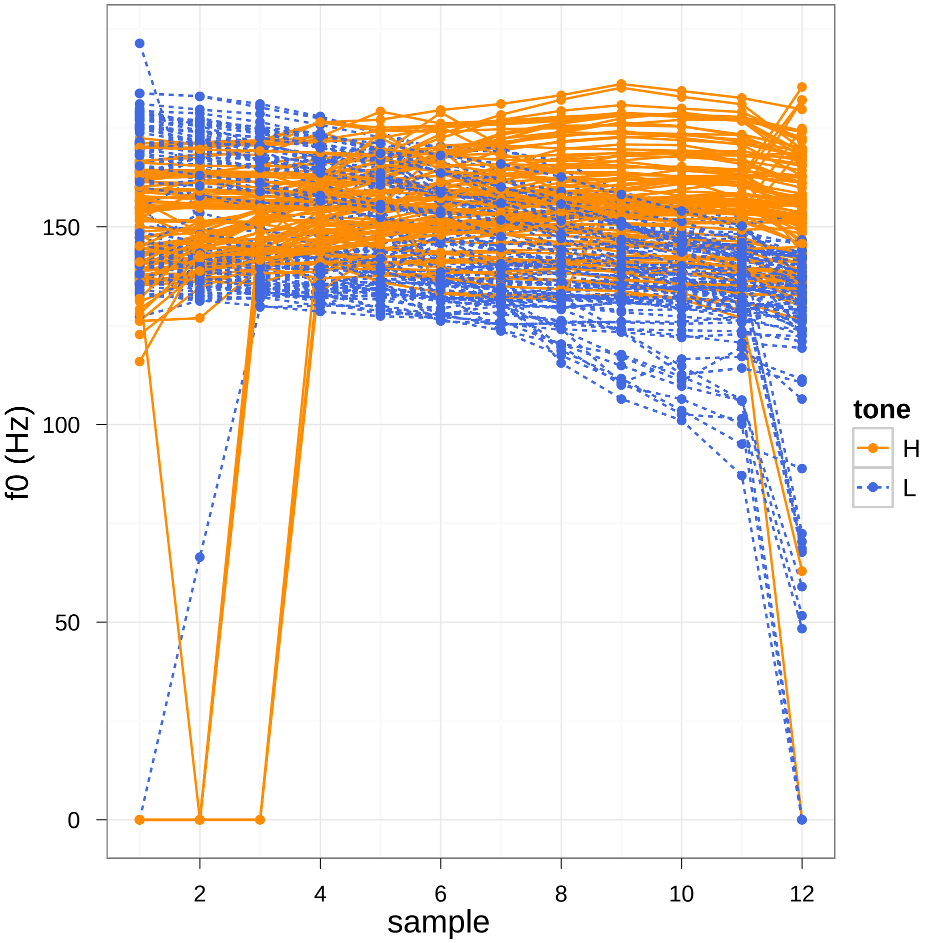


Figure 5.2: f0 contours from Bole speaker m1, with 12 samples over the syllable extracted from sentence-medial position. There is little sign of the effect of voice quality on pitch tracking, but there appears to be some edge effects causing discontinuities in the f0 tracks.

Tone	5-level system	Tone letters	Contour shape
1	55	˥	High level
2	35	˧˥	Rise
3	21(4)	˨˩˦	Low or Fall-Rise (Dipping)
4	51	˥˩	Fall
5	Toneless	Neutral/light	

Table 5.1: Labels for Mandarin tonal categories (Chao, 1968)

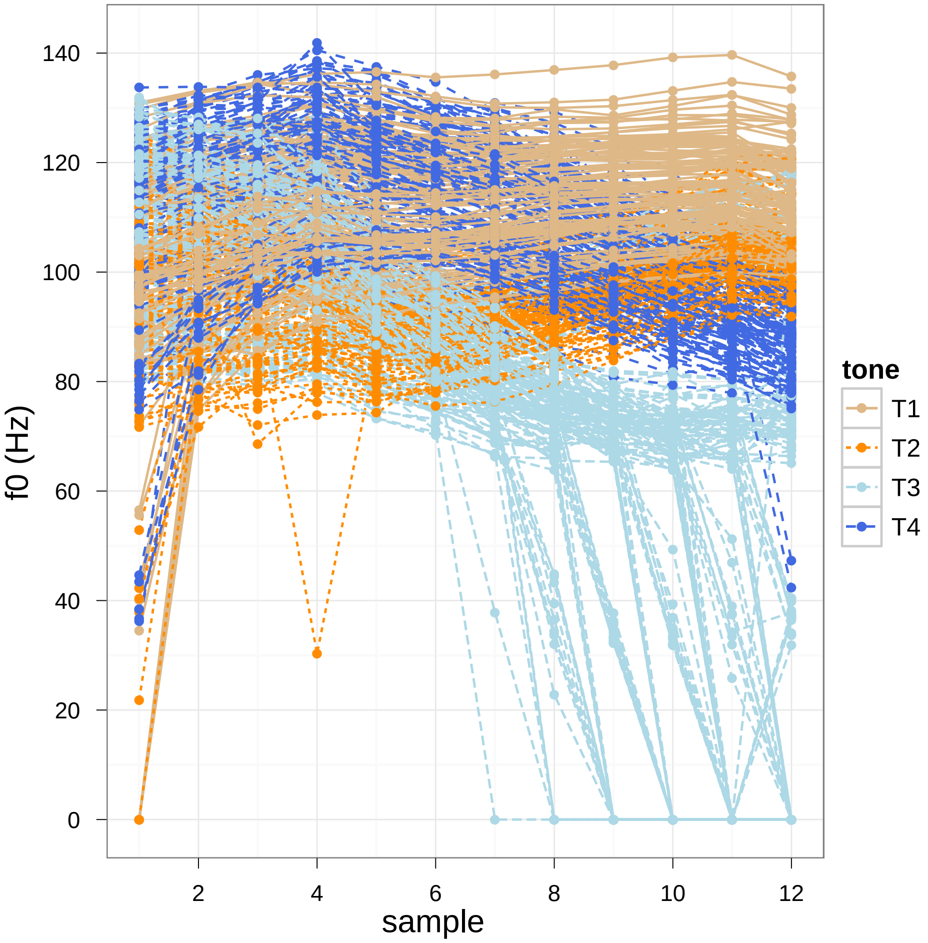


Figure 5.3: f0 contours from Mandarin speaker m1, with 12 samples over the syllable extracted from sentence-medial position. Many T3 utterances show the typical signature of pitch halving in creaky regions.

In Cantonese (Table 5.2), Figure 5.4 shows that T4 was mostly produced without creak interacting with the f0 contour, but there were at least a few exceptions to this.

Tone	5-level system	Tone letters	Contour shape
1	55	˥	High level
2	25/35	˨˥/˧˥	High rising
3	33	˧	Mid level
4	21	˨˩	Mid-low falling
5	23/13	˨˧/˩˧	Mid-low rising
6	22	˨	Mid-low level
7	5	˥	High stopped
8	3	˧	Mid stopped
9	2	˨	Mid-low stopped

Table 5.2: Labels for Cantonese tonal categories

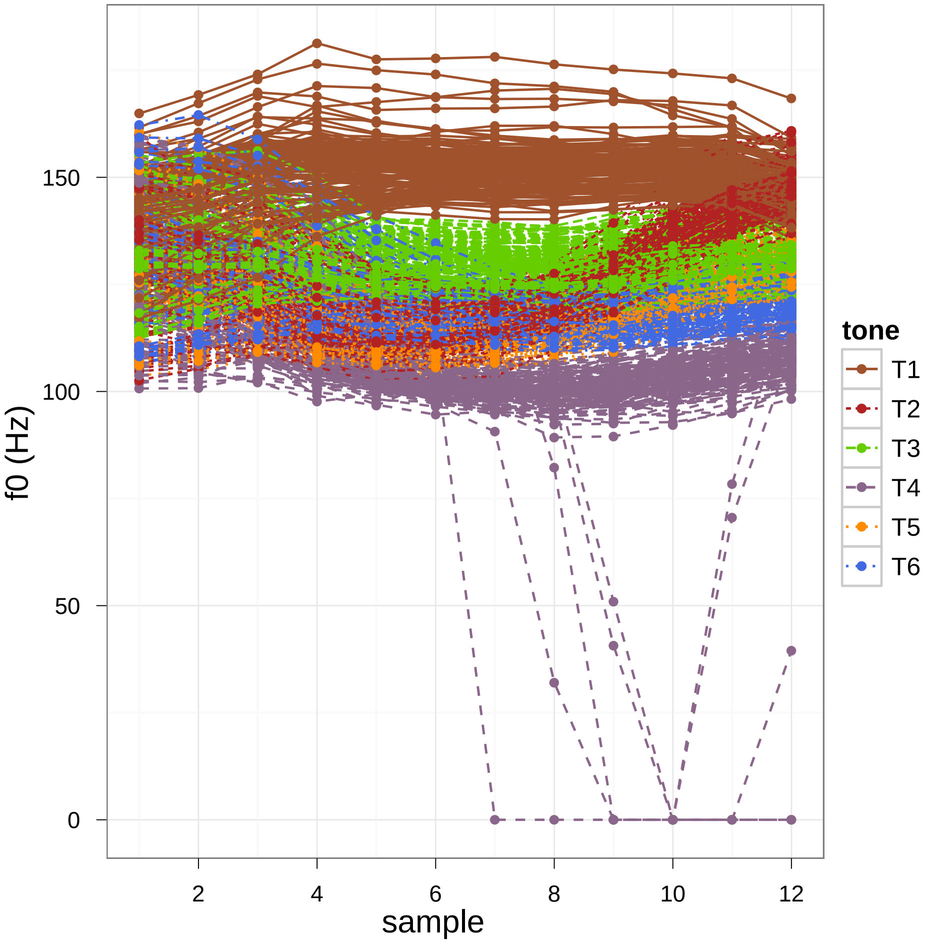


Figure 5.4: f0 contours from Cantonese speaker m1. Some T4 utterances show creak.

In Hmong (Table 5.3), Figure 5.5 shows that the m-tone was mostly produced without creak that affected the smoothness of the f0 contour, almost without exception. Also, the breathy g-tone does not seem to show f0 irregularities.

Tone	Contour shape (Esposito et al., 2009)	Ratliff (1992)	Voice quality
b-tone	High-rising	55 ˥	Modal
null-tone	Mid	33 ˧	Modal
s-tone	Low	22 ˨	Modal
j-tone	High-falling	52 ˥˨	Modal
v-tone	Mid-rising	24 ˨˦	Modal
m-tone	Low-falling	21 ˨˩	Creaky
g-tone	Mid-falling (males), High-falling (females)	42 ˦˨	Breathy

Table 5.3: Labels for Hmong tonal categories

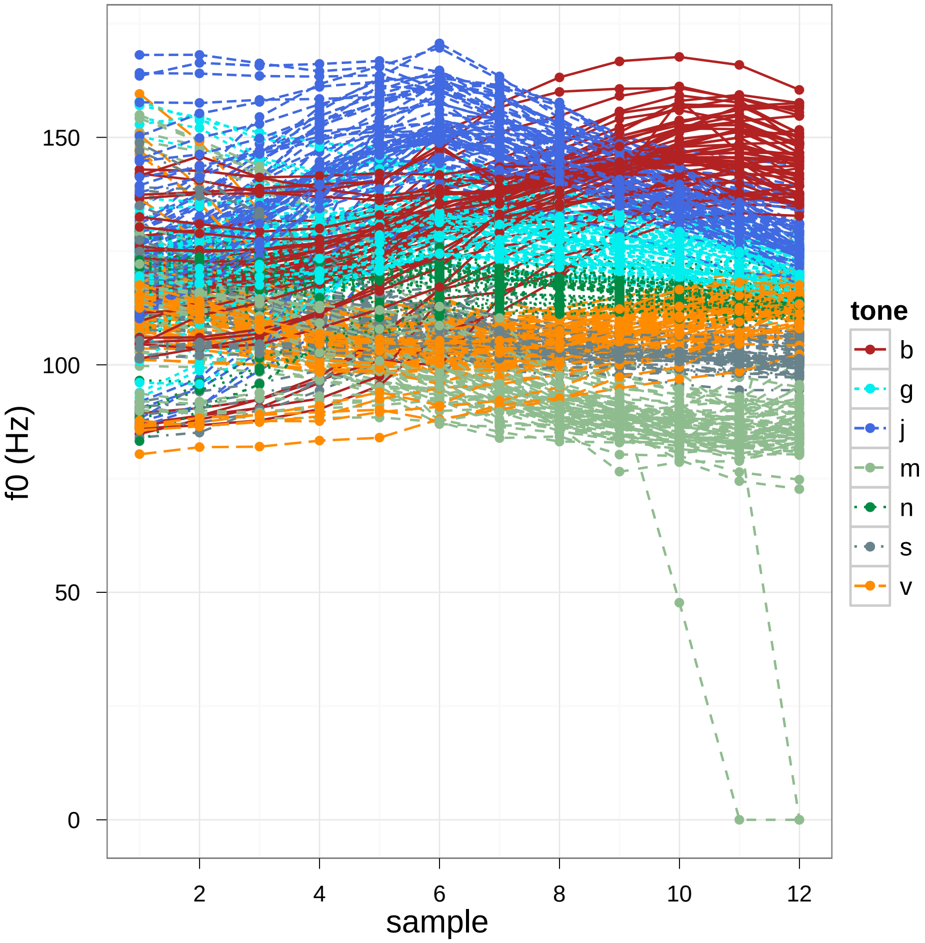


Figure 5.5: f0 contours from Hmong speaker m6, with 12 samples over the syllable extracted from sentence-medial position. The creaky m-tone shows some pitch halving, and the breathy g-tone shows a distinct f0 contour from the j-tone for this male speaker.

Overall, although we did our best to abstract away from voice quality and work within f0 value-based parameter spaces, it was not possible to do this entirely cleanly except perhaps in Bole, even for speakers hand-picked to have minimal creak in their productions. An option to force a clean abstraction for methodological purposes would be to remove f0 contours exhibiting discontinuities under some criteria from the data set, but we did not pursue that here. In the convex hull plots in the following sections, the ubiquitous presence of creak interacting with f0 is still visible in the creaky tones in each language: T3 in Mandarin, T4 in Cantonese, and the m-tone in Hmong. The convex hulls for these tones tend to be larger in area than for the other tones, due to what one might call outliers, but more accurately, data that shows interaction of voice quality with idealized f0-based parameters. Since the creaky voice cues are informative to the listener, as we showed in Chapter 3, the effect of creak on dispersing the creaky tones further from other tones within a tonal inventory in the plots is not a spurious effect of outliers.

5.4 Temporal resolution within a parameter set

Overall, we found that LDA classifiers with f0 values and/or f0 change values performed remarkably well in monosyllables extracted from connected speech in single-speaker spaces for Bole, Mandarin, Cantonese, and Hmong, with classification accuracies generally around 90% for parameter sets of cardinality greater than two. This follows the pattern in Chapters 4 and 2, in which machine classification accuracy was also high; in those chapters, we used by-speaker z-score standardization in preprocessing for speaker-independent recognition over 5 speakers, but here in single speaker spaces, the only preprocessing was smoothing f0 values (averaging over time slices) and log transforming them.

The high accuracy for LDA classification with only a few real values is impressive, but not wholly surprising, because the classification task is over single speaker spaces, with speakers with quite dispersed tones. In Ma et al. (2005), Cantonese listeners were tested on monosyllables extracted from connected speech (in a frame sentence between a T4 and a T6) from two speakers, the stimuli were blocked by speaker (serial single speaker recognition). This is a situation similar to the classification task here, except that there was more contextual variability in our stimuli, and listeners were quite accurate in tonal identification in that experiment. Listener accuracies were quite high except for the T2 rise which was perceived mostly as T5. They were, for T1-T6 respectively,86.1%, 38.9%, 63.9%, 97.2%, 99.1%, 86.1% (From Table 2 in Ma et al. (2005)).

Turning to results for the effect of temporal resolution, we found, like in the computational modeling in Chapter 4, that the separability of tonal concepts was minimally affected by temporal resolution within parameter sets of (log transformed) f0 values alone, or f0 change values alone. Tables 5.4, 5.5, 5.6, and 5.7 give the leave-one-out classification accuracies for each language, for each parameter set, for each temporal resolution tested. They show that for every language, there was a jump in accuracy from 1 to 2 samples and a smaller jump from 2 to 3 samples, but otherwise, classification accuracies are very similar across temporal resolutions, within a parameter set. We remind the reader that the blank cells in the tables are because we tested more temporal resolutions for static f0 parameter sets than the parameter sets involving f0 change values, since the f0 change values at particular temporal resolutions needed to be calculated from static f0 parameter sets at other temporal resolutions.

Parameters/samples	1	2	3	4	5	7	10	12
log f0	68.50	90.50	89.50	89.50	91.50	89.50	88.00	88.88
f0 change	82.50	78.00	80.50		75.50		80.25	
Both		91.00		88.50			89.50	

Table 5.4: Comparison of LDA leave-one-out classification accuracy across parameters and sampling resolutions for Bole speaker m1

Parameters/samples	1	2	3	4	5	7	10	12
log f0	54.50	74.84	89.39	89.18	90.21	89.45	90.37	90.69
f0 change	58.82	81.17	83.82		82.96		82.36	
Both		74.19		89.39			87.66	

Table 5.5: Comparison of LDA leave-one-out classification accuracy across parameters and sampling resolutions for Mandarin speaker m1

Parameters/samples	1	2	3	4	5	7	10	12
log f0	61.06	80.58	93.58	94.22	93.93	94.64	94.02	94.95
f0 change	45.92	76.29	77.46		79.70		79.18	
Both		81.23		92.65			93.70	

Table 5.6: Comparison of LDA leave-one-out classification accuracy across parameters and sampling resolutions for Cantonese speaker m1

Parameters/samples	1	2	3	4	5	7	10	12
log f0	57.04	80.46	94.08	92.61	93.00	93.55	92.83	92.91
f0 change	47.78	75.40	75.87		75.71		73.30	
Both		81.92		92.83			92.92	

Table 5.7: Comparison of LDA leave-one-out classification accuracy across parameters and sampling resolutions for Hmong speaker m6

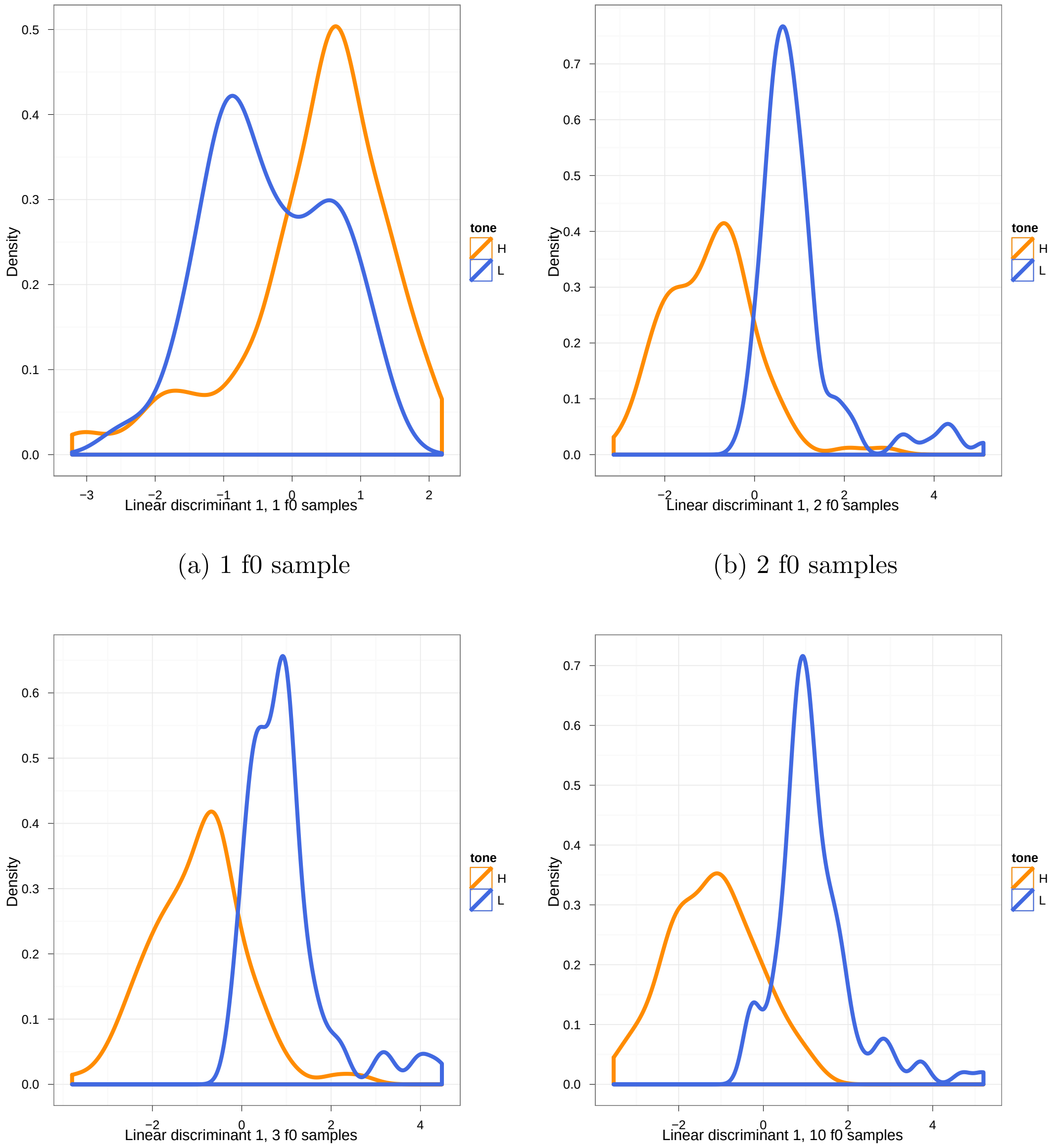
We show some highlights from these results with density plots and convex hull plots. Figure 5.6 displays the effect of increasing temporal resolution for f0 values in Bole, a language with only a contrast between H and L tones. While there is an increase in separability from 1 (Fig. 5.6a) to 2 samples (Fig. 5.6b),[1]. separability is approximately constant for higher resolutions, with 3 (Fig. 5.6c) or 10 samples (Fig. 5.6d). This was the general trend in all languages, for both f0 values and f0 change values.

Figure 5.7 displays the different effects between f0 values and f0 change values of increasing temporal resolution from 2 to 3 samples. For 2 f0 values (Figs. 5.7a, 5.7b), the convex hulls of T1, T2, and T3 show significant overlap, but the overlap is much reduced with 3 samples. However, there is litte difference in separability between 2 and 3 values for f0 change (Figs. 5.7c, 5.7d).

As discussed in Chapter 4, there have been multiple suggestions in the literature that a minimum of 3 f0 values is necessary to capture the turning point difference in contours between the T2 rise and T3 fall-rise in isolation. For monosyllables extracted from connected speech, the large increase in separability from 2 to 3 samples in a f0-value parameter space is consistent with the idea of landmark/target f0 values for delineating the contour shape. Thus, it is noteworthy that there is no marked increase in separability from 2 to 3 samples in a space defined over f0 *change* rather than f0 values: classification accuracy was 81.17% with 2 samples and 83.82% with three. Since the f0 change values are calculated between f0 values, there is more information in 2 f0 change values (as linear combinations of 2 pairs of 2 f0 values each) than 2 f0 values for describing contour shape. If 3 f0 values are sufficient for maximizing separability, as evidenced in Table 5.5 where classification accuracy asymptotes to a ceiling for 3 samples

[1] But see §5.6 for a discussion of taking a sample over a narrower window of time, rather than taking the mean f0

or more, then it is logical that 2 f0 change values, e.g. a negative value (fall) followed by a positive one (rise) are sufficient as well, with little discriminatory power drawn from higher temporal resolution. Indeed, classification accuracy asymptotes in Table 5.5 for f0 change for 2 f0 samples or higher.



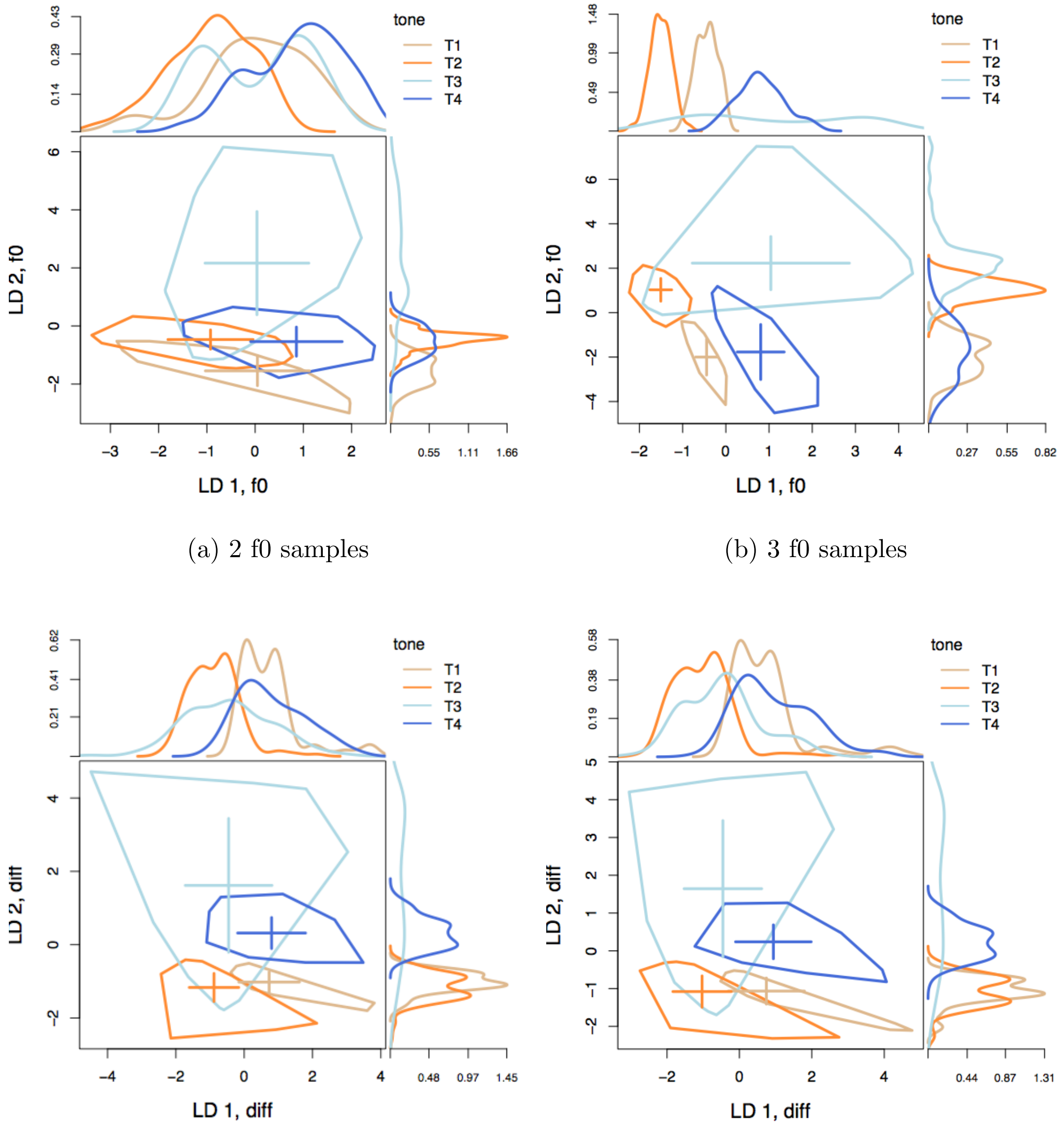
(a) 1 f0 sample

(b) 2 f0 samples

(c) 3 f0 samples

(d) 10 f0 samples

Figure 5.6: Density plots in LDA-optimized space for Bole speaker m1, for 1, 2, and 3 log-transformed f0 values. There is much less overlap between the H and L tone categories from 1 (a) to 2 samples (b), but little difference between overlap for 2 to 3 (c) to 10 samples (d).



(a) 2 f0 samples

(b) 3 f0 samples

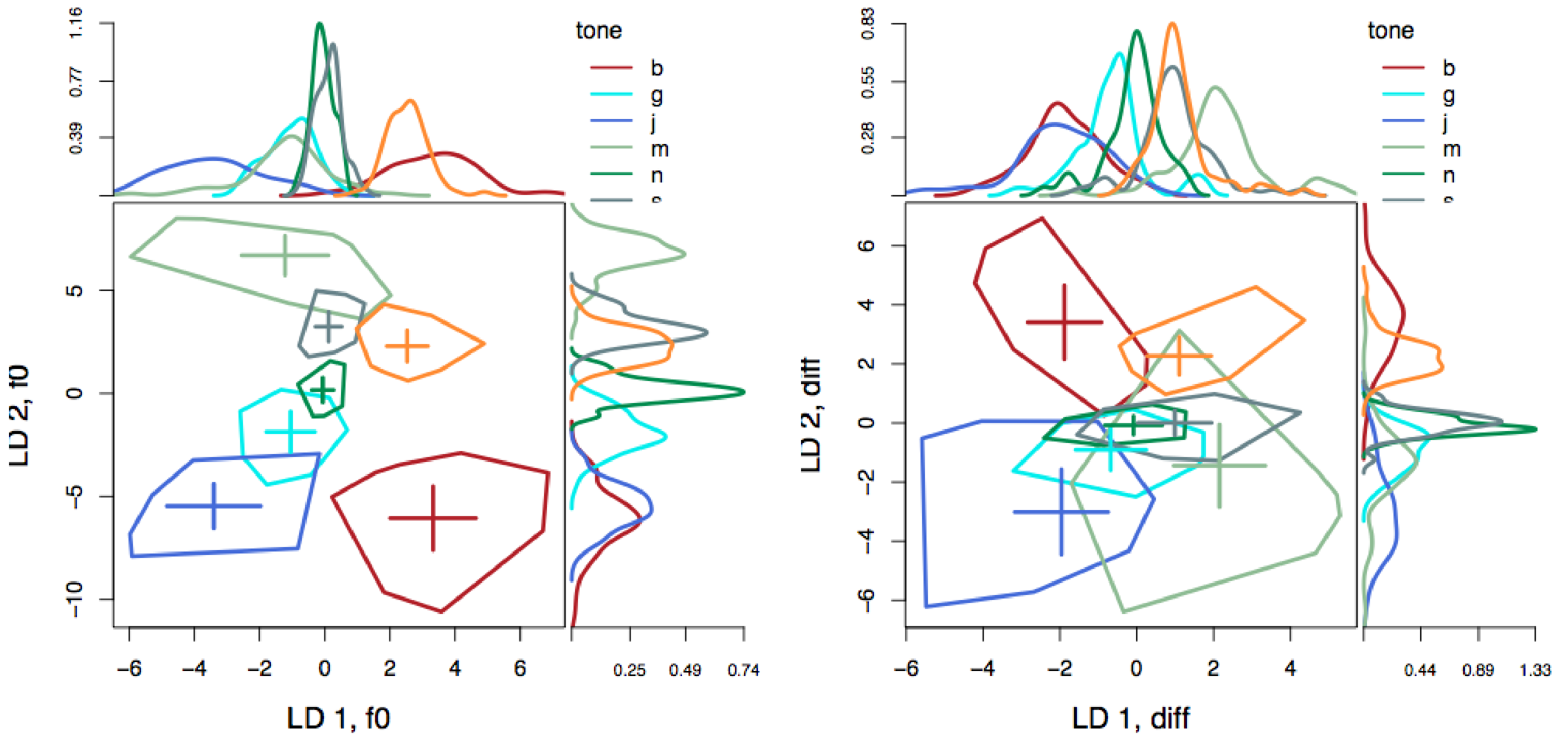
(c) 2 f0 change samples

(d) 3 f0 change samples

Figure 5.7: Convex hull plots in LDA-optimized 2-D spaces for Mandarin speaker m1, for 2 and 3 log-transformed f0 values and f0 change values. Top: there is much less overlap between T1, T2, and T3 for 3 samples (b) than 2 sample of log(f0) (a). Bottom: but there is little difference in separability between 2 (c) and 3 samples for f0 change (d).

5.5 The parameterization of f0: f0 and f0 change

We could not generalize Gauthier et al. (2007)'s finding that Mandarin tones across multiple speakers are better categorized in some sense with dynamic f0/velocity parameters than static f0 parameters. Dynamic f0 parameter spaces did not produce consistently greater separability in single speaker spaces in Mandarin or in the other languages, which all had level tone contrasts. For a given temporal resolution, classification accuracy was higher in a parameter space defined over static f0 values than in one defined over dynamic f0 values for every language in every temporal resolution compared, with two exceptions, for the 1 sample comparison between f0 and f0 change in Bole, and the 2 sample comparison in Mandarin. The overwhelming pattern, though, was that static f0 spaces were better able to separate out tones, as illustrated in Figure 5.8. Even with 10 samples over the syllable, f0 change parameters could not separate the Hmong tones well (Figure 5.8b), but a parameter space defined by 10 f0 samples resulted in minimal overlap between convex hulls (Figure 5.8a).



(a) 10 f0 samples

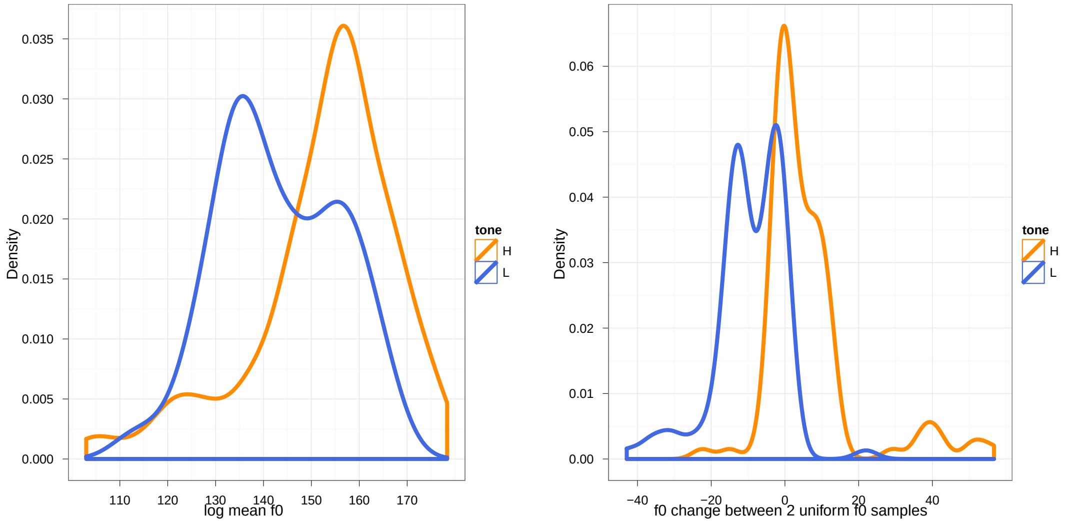
(b) 10 f0 change samples

Figure 5.8: Convex hull plots in LDA-optimized 2-D spaces for Hmong speaker m6, for 10 log-transformed f0 values (a) and for 10 f0 change values (b). Separability is much higher in the f0 value space. In (b), the m, n, and s tones (all level tones) nearly fully overlap. However, convex hulls for all tones in (a) show little overlap.

The exception is a surprise. Table 5.4 shows that Bole classification accuracy was 68.50% with just mean log f0, but it was 82.50% with mean slope/f0 change. Figure 5.9b shows that although both H and L tones had little overall change in f0 over the syllable, the majority of L tones had a negative mean slope—there is even a bimodal distribution, which picks out a subset of L tones with negative f0 slope, and a subset of L tones with approximately constant f0. This is due to allophonic variation: in Bole, L tones following H tones fall, but L tones following L tones are more flat. Dynamic f0 parameters were better able to segregate out L tones following H tones than following L tones. Figure 5.10 shows f0 contours for H and L tones in either the first tone of the bitone (top) or the second (bottom). For the L tone in both positions, there is a clear bifurcation between flatter contours and ones falling from previous H tones. The corresponding context for H tones, preceded by Ls, doesn't show as strong a bifurcation, and this is reflected in the lack of a bimodal distribution for H tones in the dynamic parameter space.

The superiority of a single dynamic parameter over a single static one in a single speaker space is quite surprising, given that Bole is a tone language with two level tones, and that level tones are often defined as single f0 targets (Goldsmith, 1976). Gauthier et al. (2007) attributed the better performance of dynamic parameters in Mandarin multispeaker spaces to the effect of speaker normalization that occurs with f0 change parametrization which is insensitive to global shifts in pitch range that occur across speakers. Here, we see that at the same temporal resolution, f0 change can better able to separate a single level tone contrast in a single speaker space.

It is worth noting though, that while our tonal production corpora uniformly sampled over all licit bitone combinations—with half of the combinations, LH and HL, creating environments for rises and falls between the two tones in the bitone and from the syllables preceding and following the bitone—different distributions of bitone combinations in input to the learner and listener could change separability characteristics; in Bayesian terms, while we assumed uniform priors over tone unigrams and bigrams, the situation for the learner/listener might be quite different.



(a) mean f0

(b) mean f0 slope

Figure 5.9: Density plots in LDA-optimized space for Bole speaker m1, for mean f0 (Hz), (a), or mean f0 slope (Hz), (b). The overlap in the H and L tone classes is smaller in the mean f0 slope parameter space. The bimodal distribution in that space is due to allophonic variation.

The greater separability of H and L level tones in dynamic f0 spaces in Bole, a 2-tone system, relative to the separability of level tones in the 6- and 7-tone systems of Cantonese and Hmong, e.g. Fig. 5.8, suggests that level tones in large tonal inventories with contour tones might be realized with more level contours than in small tonal inventories with only level tones. We note, though, that in recording the languages with larger tonal inventories (Cantonese, Hmong), we

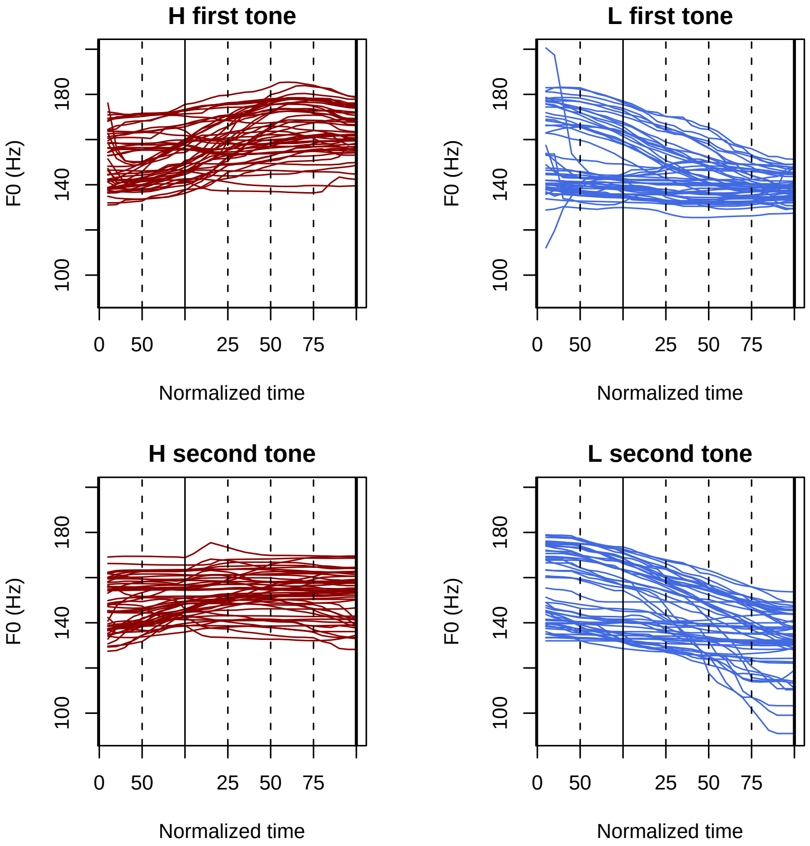


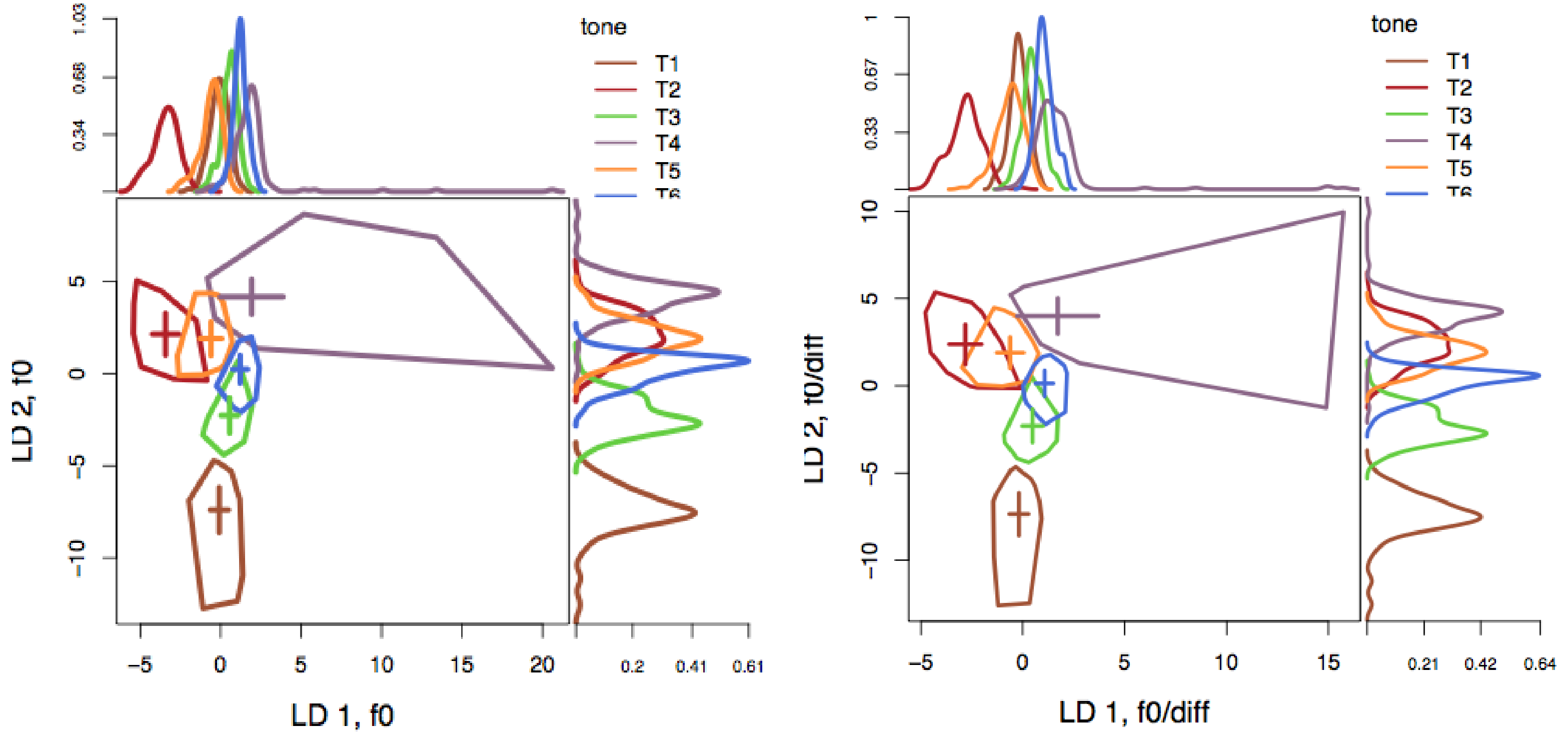
Figure 5.10: Allophonic variation in Bole speaker m1's productions, showing f0 contours for H and L tones in the first and second members of the elicited bitones. The time scale shows percentage of the segment, and there are two segments, an onset consonant and a nucleus vowel, separated by a solid line. The striking bifurcation in the realization of the L tones is due to the fall from a preceding H vs. a flatter contour following a L. The larger fall in some of the Ls realized as the second member of the bitone is due to a sentence frame in which the L was sentence-final.

fixed the syllables preceding and following the target bitones to be mid-level tones, unlike in Bole, where they varied over both H and L tones, promoting more dynamic f0 contours. Still, in the Cantonese temporal domain experiment

(Chapter 2), confusability in the isolation context came from two separate sources: level tones were confused with level tones and contour tones with contour tones. To separate level tones and contour tones in languages like Cantonese and Hmong, then, a natural idea is to define spaces over both static and dynamic parameters.

In a parameter space of fixed size, we found that equally dividing the parameter set between static and dynamic f0 parameters yielded tonal separability that was comparable to the highest performing parameter set between all static or all dynamic f0 parameters, with the exception of 2-parameter spaces in Mandarin, in which classification accuracy was around 74-75% for one static and one dynamic parameter, as well as for two static parameters, slightly lower than the 81.17% accuracy for two dynamic parameters. Figures 5.7a and 5.7c show that separability between the two Mandarin parameter spaces with 2 static or dynamic parameters is similar. The main contribution to increased separability with 2 f0 change values relative to 2 f0 values is that the T2 rise is better separated from the T3 low/fall-rise and the high level tone T1; it makes sense that dynamic information is more informative for discriminating these tones with similar f0 levels.

The nearly exceptionless majority pattern, though, is exhibited in Figure 5.11 for Cantonese 4-parameter spaces. The plot for a f0 value space and the plot for one with f0 values and f0 change values are very similar, and classification accuracy was 94.22% and 92.65% for the all f0 and mixed parameter set, respectively. The most overlapped tone classes were the rises T2/T5 and the mid and mid-low level tones T3/T6; both pairs are known to be merged in some Cantonese speakers/listeners.



(a) 4 f0 samples

(b) 2 f0 samples, 2 f0 change samples

Figure 5.11: Convex hull plots in LDA-optimized 2-D spaces for Cantonese speaker m1, for 4 log-transformed f0 values (a), compared to 2 log-transformed f0 values and 2 f0 change values (b). The overlaps between convex hulls is similar in both plots. The T2 and T5 rises show overlap, as do the close level tones T3 and T6.

5.6 Tonal separability with samples over narrow windows of time near the syllable onset and offset

Throughout this chapter, we extracted sample parameters as averages, e.g. when we extracted 5 f0 samples, we extracted mean f0 samples from each of five evenly divided windows over the syllable. This is a common practice in automatic tonal recognition, e.g. Qian et al. (2007). Especially in cases of low sampling resolution, though, the smoothing effect of this averaging might result in lower tonal separability than if samples taken over finer time windows were taken. In exploratory work for future directions, we checked tonal separability with 1 sample taken near the syllable onset and near the syllable offset for both f0 and f0 change values. For both the f0 and f0 change values, we sampled the 2nd and 4th values extracted for the 5-sample condition for each language, as shown below in Tables 5.8, 5.9,5.10, and 5.11.

Sampling/Parameter	log f0	f0 change
Average	68.50	82.50
2nd/5 samples (onset)	57.00	84.50
4th/5 samples (offset)	79.00	71.00

Table 5.8: Comparison of LDA leave-one-out classification accuracy across parameters for 1-sample extraction from different temporal windows for Bole speaker m1: the average value over the syllable and the 2nd and 4th log f0 or f0 change value out of 5 samples over the syllable.

In all but one case, LDA leave-one-out classification accuracy was higher with at least one of the samples averaged over 1/5 of the syllable duration than with a sample averaged over the entirety of the syllable duration. In all but one case (for

Sampling/Parameter	log f0	f0 change
Average	54.50	58.82
2nd/5 samples (onset)	44.26	42.22
4th/5 samples (offset)	65.80	77.95

Table 5.9: Comparison of LDA leave-one-out classification accuracy across parameters for 1-sample extraction from different temporal windows for Mandarin speaker m1: the average value over the syllable and the 2nd and 4th log f0 or f0 change value out of 5 samples over the syllable.

Sampling/Parameter	log f0	f0 change
Average	61.06	45.92
2nd/5 samples (onset)	54.98	42.75
4th/5 samples (offset)	66.46	49.17

Table 5.10: Comparison of LDA leave-one-out classification accuracy across parameters for 1-sample extraction from different temporal windows for Cantonese speaker m1: the average value over the syllable and the 2nd and 4th log f0 or f0 change value out of 5 samples over the syllable.

f0 change in Bole), the sample taken at the offset (the 4th of 5 samples over the syllable) yielded higher LDA accuracy than the sample taken at the onset (the 2nd of 5 samples over the syllable). The difference in accuracy was sometimes large: in Mandarin, classification accuracy was 20 to 30% higher with the offset than the onset sample. The increased tonal separability with acoustic information near the offset of the syllable is consistent with Khouw and Ciocca (2007)'s result showing that Cantonese tones in isolation were maximally separable near the syllable offset and with studies of tonal coarticulation showing that carryover

Sampling/Parameter	log f0	f0 change
Average	57.04	47.78
2nd/5 samples (onset)	48.33	46.14
4th/5 samples (offset)	70.22	53.11

Table 5.11: Comparison of LDA leave-one-out classification accuracy across parameters for 1-sample extraction from different temporal windows for Hmong speaker m6: the average value over the syllable and the 2nd and 4th log f0 or f0 change value out of 5 samples over the syllable.

coarticulation is stronger than anticipatory coarticulation, as discussed in §4.3.3 in Chapter 4 on page 167.

While these exploratory analyses suggest that tonal separability in a space defined by a single real value is higher if the real value is extracted over a smaller temporal window over the syllable than the entirety of the syllable, the highest classification accuracy achieved with a single sample, even extracted over the smaller temporal windows, was comparable or up to around 10% lower than accuracies achieved with 2 samples (as means over the first and second halves of the syllable). Thus, our results showing the insensitivity of tonal separability to sampling resolution are not contradicted by the analyses described in this section.

5.7 General discussion

Up to this chapter, we have discussed only results based on work in Cantonese. We chose to perform all our tonal perception experiments in Cantonese since it has level tone contrasts, and thus is more representative of the large majority of the world's tone languages which have level tone contrasts (Maddieson, 1978) than Mandarin, the most commonly used language for studying tonal perception. But as long as one bases ideas about tonal representation on any single language, those ideas will be biased by that language. As we discussed in Chapter 1, to study how tones are defined with the larger goal of understanding how tonal concepts are learned by humans, it is necessary to analyze cross-linguistic data.

By studying f0-based tonal spaces of single speakers in Bole, Cantonese, Mandarin, and Hmong with computational modeling in this chapter, we were able to explore how results from our experiments and computational modeling in Cantonese and Gauthier et al. (2007)'s modeling work in Mandarin generalized to other languages.

First, we found that despite attempts to abstract away from voice quality information in the speech signal and to work within f0-based spaces, we were not able to cleanly hack f0 parameters away from the larger body of voice quality parameters. While breathy phonation in the signal in Hmong for the speaker used here didn't seem to interact much with the idealized f0 values extracted, the presence of creak was evident across all languages except Bole. Given Maddieson (1977, 1978)'s idea that additional dimensions are introduced for tonal inventories that are larger in size, the evidence here for little use of voice quality parameters other than f0 values in Bole is of interest (This was true in the other 4 speakers recorded as well).

With a tonal inventory of size 2, Bole has the smallest tone inventory that a tone language can have. In Maddieson (1977, 1978)'s implicational hierarchy, if a language has contour tones it also has level tones, and if a language has complex contours (with more than one extrema, such as in a fall-rise, as opposed to a fall or a rise) it also has simple contours (falls and rises). Might we add, that if voice quality cues beyond pitch cues are informative in a tone language, then the language also has contour tones, or the language has at least n tones? For languages with contrastive phonation like Hmong and Vietnamese (Chapter 3), these seems like a reasonable statement, as Hmong and Vietnamese have tonal inventories of large size, and multiple contour tones (thought not clearly "complex" contour tones in Maddieson's sense). However, what about register languages, as discussed in Chapter 3? Most have two, such as Burmese/Khmer, and if more studies show that in these languages, both pitch and other voice quality cues non-independently determine human tonal perception, then this hypothetical voice quality implicational universal cannot be correct.

We also found that the separability of tones in the sample of languages and speakers studied here was extremely high for temporal resolutions beyond 2 samples over the syllable: leave-one-out classification accuracies from the LDA classifiers were around 90% and the convex hull plots showed near linear separability in each language in simple f0 value and/or f0 change parameter spaces. Because we were working with single speaker spaces of individuals we picked for having distinct tonal contrasts, this result is not too surprising. But is also not unsurprising: there was no contextual information available to the classifier, unlike in the computational modeling of the effect of having f0 information from neighboring syllables in Chapter 2; yet classification accuracy for syllables extracted from varying bitone contexts was near-perfect across languages with as few as 3 f0 values. We think it is too early, without more analysis from other speakers, to make much of the raw classification accuracies here. We don't think it would be appropriate to generalize to tonal perception over single speakers cross-linguistically. However, our exploratory study is convincing in showing that extremely sparsely tonal spaces—in the type of parameter (just based on f0 values) and the resolution of parameter (only 3 values)—can be sufficient for representing the tones from some speakers.

The *relative* accuracies across temporal resolutions show evidence that our hypothesis that coarse temporal resolution is sufficient for good separability of tonal categories (Chapter 4) generalizes to other tone languages than Cantonese. Accuracy reached a maximum by a temporal resolution of 3 samples over the syllable and was stable for higher resolutions across languages, and the convex hull plots show near linear separability as well. Thus, low dimensionality in the temporal resolution of tonal representations appears to be something quite general about tones and not language-specific. This is evidence for something basic, common to humans, that causes this. One obvious constraint on temporal resolution comes from the generating source, from articulatory constraints on the speed of f0 change (Ohala and Ewan, 1973; Sundberg, 1979; Xu and Sun, 2002). Since it is not clear exactly how f0 is controlled in production, though, it is hard to model what this constraint is explicitly.

The comparison of separability with the static and/or f0 values showed that for a fixed size of the parameter set, static f0 values generally yielded higher separability than f0 change values, with an exception for parameter sets of cardinality 1 in Bole and of cardinality 2 in Mandarin. Thus, it appears that unlike the low dimensionality in the temporal representation of tones, the existence of a single invariant type of parameter, such as f0 values or f0 change values, for representing tones is not something that is common across tone languages. Since f0 and f0 change values are linearly dependent anyway, it seems difficult to motivate an auditory basis for one type of parameter leading to greater tonal separability than the other. In any case, we found that by combining static and dynamic f0 values, separability for a parameter space of a given size was nearly exceptionlessly as high as separability with static or dynamic f0 values alone.

In summary, even without access to directly relevant human perceptual data, we were able to use computational modeling in a range of tone languages—Bole, Mandarin, Cantonese, and Hmong—to better understand how tones are represented in human cognition. We were able to generalize two results from Cantonese: voice quality interactions with f0 contours were present, and coarse temporal resolution in simple f0 value parameter sets were sufficient for near linear separability of tones in every language.

Bibliography

Abramson, Arthur S. 1972. Tonal experiments with whispered Thai. In *Papers on linguistics and phonetics in memory of Pierre Delattre*, edited by A. Valdman. Mouton, The Hague, pages 31–44.

Abramson, Arthur S., Theraphan L-Thongkum, and Patrick W. Nye. 2004. Voice Register in Suai (Kuai): An Analysis of Perceptual and Acoustic Datauai (Kuai): An Analysis of Perceptual and Acoustic Data. *Phonetica* 61(2-3):147–171.

Abramson, Arthur S. and Theraphan Luangthongkum. 2009. A fuzzy boundary between tone languages and voice-register languages. In *Frontiers in phonetics and speech science*, edited by G. Fant, H. Fujisaki, and J. Shen. The Commercial Press, pages 149–155.

Abramson, Arthur S., Patrick W. Nye, and Theraphan Luangthongkum. 2007. Voice Register in Khmu': Experiments in Production and Perception. *Phonetica* 64(2-3):80–104.

Andruski, Jean E. 2006. Tone clarity in mixed pitch/phonation-type tones. *Journal of Phonetics* 34(3):388–404.

Andruski, Jean E. and James Costello. 2004. Using Polynomial Equations to Model Pitch Contour Shape in Lexical Tones: An Example from Green Mong. *Journal of the International Phonetic Association* 34(02):125–140.

Andruski, Jean E. and Martha Ratliff. 2000. Phonation Types in Production of Phonological Tone: The Case of Green Mong. *Journal of the International Phonetic Association* 30(1-2):37–61.

Assmann, Peter and Quentin Summerfield. 2004. The perception of speech under adverse conditions. In *Springer Handbook of Auditory Research*, volume 18. Springer-Verlag, New York, pages 231–308.

Baayen, R. H. 2008. *Analyzing linguistic data: a practical introduction to statistics*. Cambridge University Press.

Baayen, R.H., D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4):390–412.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72(1):32–68.

Barry, Johanna G. and Peter J. Blamey. 2004. The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. *The Journal of the Acoustical Society of America* 116(3):1739–1748.

Bashford, James A., Keri R. Riener, and Richard M. Warren. 1992. Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and psychophysics* 51(3):211–217.

Bashford, James A., Richard M. Warren, and Christopher A. Brown. 1996. Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. *Perception & Psychophysics* 58(3):342–350.

Bates, Douglas and Martin Maechler. 2010. *lme4: Linear mixed-effects models using S4 classes*. URL `http://lme4.r-forge.r-project.org/`, R package version 0.999375-37.

Belotel-Grenie, Agnés and Michel Grenie. 1994. Phonation types analysis in Standard Chinese. In *ICSLP-1994*. pages 343–346.

Belotel-Grenié, Agnés and Michel Grenié. 1997. Types de phonation et tons en chinois standard. *Cahiers de Linguistique - Asie Orientale* 26(2):249–279.

Belotel-Grenié, Agnès and Michel Grenié. 2004. The Creaky Voice Phonation And The Organisation Of Chinese Discourse. In *TAL-2004*. pages 5–8. URL `http://www.isca-speech.org/archive/tal2004/tal4\_005.html`.

Bennett, Kristin P. and Erin J. Bredensteiner. 2000. Duality and Geometry in SVM Classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pages 57–64.

Best, Catherine T. 2003. Peeling back the layers of time: integrating speech perception on the scales of stimulus time, experiential time, and developmental time. *Journal of Phonetics* 31(3-4):613–618.

Blankenship, Barbara. 2002. The timing of nonmodal phonation in vowels. *Journal of Phonetics* 30(2):163–191.

Blum, Lenore. 2004. Computing over the reals: where Turing meets Newton. *Notices of the American Mathematical Society* 51(9):1024–1034.

Blum, Lenore, Felipe Cucker, Michael Shub, and Steve Smale. 1997. *Complexity and real computation*. Springer.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery* 36(4):929–965.

Boersma, Paul and David Weenink. 2010. Praat: doing phonetics by computer (Version 5.1.32) [Computer program]. `http://www.praat.org`.

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge University Press.

Brainard, D. H. 1997. The psychophysics toolbox. *Spatial vision* 10:443–446.

Brunelle, Marc. 2009. Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics* 37(1):79–96.

Burges, Christopher J.C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2(2):121–167.

Chang, Chih-Chung and Chih-Jen Lin. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chao, Yuen-Ren. 1930. A system of tone-letters. *Le Maître Phonétique* 45:24–27.

Chao, Yuen Ren. 1956. Tone, intonation, singsong, chanting, recitative, tonal composition and atonal composition in Chinese. In *For Roman Jakobson: essays on the occasion of his sixtieth birthday*, edited by Morris Halle, Horace Lunt, Hugh McLean, and Cornelis van Schooneveld. Mouton, pages 52–59.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. University of California Press, Berkeley, CA.

Chen, Matthew Y. 2000. *Tone sandhi*. Cambridge University Press.

Chen, Sim-Horng and Yih-Ru Wang. 1995. Tone recognition of continuous Mandarin speech based on neural networks. *Speech and Audio Processing, IEEE Transactions on* 3(2):146–150.

Ciocca, Valter and Albert S. Bregman. 1987. Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and psychophysics* 42(5):476–484.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12(4):335–359.

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2001. *Introduction to algorithms*. MIT Press, 2nd edition.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Cover, Thomas M. and Joy A. Thomas. 2006. *Elements of information theory*. Wiley-Interscience, 2nd edition.

Da, Jun. 2004. A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. The studies on the theory and methodology of the digitized Chinese teaching to foreigners. In *Proceedings of the 4th International Conference on New Technologies in Teaching and Learning Chinese*, edited by Pu Zhang, Tianwei Xie, and Juan Xu. pages 501–511.

Dannenbring, Gary L. 1976. Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology* 30(2):99–114.

Davison, Deborah S. 1991. An acoustic study of so-called creaky voice in Tianjin Mandarin. *Working Papers in Phonetics, Department of Linguistics, UCLA* 78:50–57.

de Boer, Bart and Patricia K. Kuhl. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4(4):129–134.

De Looze, Céline and Stéphane Rauzy. 2009. Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. In *INTERSPEECH-2009*. pages 2919–2922.

Deshmukh, O., C.Y. Espy-Wilson, A. Salomon, and J. Singh. 2005. Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech. *Speech and Audio Processing, IEEE Transactions on* 13(5):776–786.

DiCanio, Christian T. 2009. The Phonetics of Register in Takhian Thong Chong. *Journal of the International Phonetic Association* 39(02):162–188.

Dillon, Brian, Ewan Dunbar, and William Idsardi. Unpublished. A single stage approach to learning phonological categories: insights from Inuktitut .

Ọdẹ́lọbí, Ọdẹ́túnjí Àjàdí. 2008. Recognition of Tones in Yorùbá Speech: Experiments With Artificial Neural Networksorùbá Speech: Experiments With Artificial Neural Networks. In *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*. Springer, Berlin, pages 23–47.

Donohue, Cathryn. 2011. The significance of 'secondary cues' for tonal identification in Fuzhou. In *Proceedings of ICPhS XVII*. pages 607–610.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. *Pattern classification*. John Wiley & Sons, Inc., 2nd edition.

Dyson, Freeman. 2004. A meeting with Enrico Fermi. *Nature* 427(6972):297.

Esling, John H. 2005. There are no back vowels: the laryngeal articulator model. *The Canadian Journal of Linguistics* :13–44.

Esposito, Christina M. 2003. Santa Ana Del Valle Zapotec Phonation. Master's thesis, University of California Los Angeles.

Esposito, Christina M., Joseph Ptacek, and Sherrie Yang. 2009. An acoustic and electroglottographic study of White Hmong phonation. *The Journal of the Acoustical Society of America* 126(4):2223.

Evanini, Keelan and Catherine Lai. 2010. The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America* 128(4):2291.

Fant, Gunnar. 1960. *Acoustic theory of speech production*. Mouton, The Hague.

Feldman, Naomi H., Thomas L. Griffiths, and James L. Morgan. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Fields, Stanley and Mark Johnston. 2005. Whither Model Organism Research? *Science* 307(5717):1885–1886.

Flynn, Choi-Yeung-Chang. 2003. *Intonation in Cantonese*. Lincom Europa.

Fok, C.Y.Y. 1974. *A perceptual study of tones in Cantonese*. Number 18 in Occasional Papers and Monographs. University of Hong Kong, Centre of Asian Studies.

Francis, Alexander L., Valter Ciocca, and Brenda Kei Chit Ng. 2006. On the (non)categorical perception of lexical tones. *Perception & Psychophysics* 65(7):1029–1044.

Gandour, Jack. 1981. Perceptual dimensions of tone: evidence from Cantonese. *Journal of Chinese Linguistics* 9:20–36.

Gandour, Jack. 1983. Tone perception in Far Eastern languages. *Journal of Phonetics* 11:149–175.

Gandour, Jack, Siripong Potisuk, Sumalee Dechonkit, and Suvit Ponglorpisit. 1992. Tonal coarticulation in Thai disyllabic utterances: a preliminary study. *Linguistics of the Tibeto-Burman area* 15(1).

Gandour, Jackson T. and Richard A. Harshman. 1978. Crosslanguage differences in tone perception: a multidimensional scaling investigation. *Language and Speech* 21(1):1–33.

Garellek, Marc and Patricia Keating. 2011. The Acoustic Consequences of Phonation and Tone Interactions in Jalapa Mazatecazatec. *Journal of the International Phonetic Association* 41(02):185–205.

Gauthier, Bruno, Rushen Shi, and Yi Xu. 2007. Learning phonetic categories by tracking movements. *Cognition* 103:80–106.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gerratt, Bruce R. and Jody Kreiman. 2001. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* 29(4):365–381.

Goldsmith, John Anton. 1976. Autosegmental phonology. Ph.D. thesis, Massachusetts Institute of Technology.

Gottfried, T. L. and T. L. Suiter. 1997. Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics* 25(2):207–231.

Gårding, Eva, Paul Kratochvil, and Jan-Olof Svantesson. 1986. Tone 4 and Tone 3 discrimination in modern Standard Chinese. *Language and Speech* 29(3):281–293.

Grassi, Massimo and Alessandro Soranzo. 2009. MLP: A MATLAB toolbox for rapid and reliable auditory threshold estimation. *Behavior Research Methods* 41(1):20–28.

Greenberg, Steven and Eric Zee. 1977. On the perception of contour tones. *UCLA Working Papers in Phonetics* 45:150–159.

Gruber, James. 2011. Perceptual Cues to Lexical Tone in Burmese. *LSA Annual Meeting Extended Abstracts* 0(1). URL `http://www.elanguage.net/journals/index.php/lsameeting/article/view/1502`.

Han, Mieko and Kong-On Kim. 1974. Phonetic variation of Vietnamese tones in disyllabic utterances. *Journal of Phonetics* 22(4):477–492.

Hant, James J. and Abeer Alwan. 2003. A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise. *Speech Communication* 40(3):291–313.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning*. Springer, second edition.

Hayes, Bruce. 2008. *Introductory Phonology*. Wiley-Blackwell.

Henderson, Eugénie J. A. 1952. The Main Features of Cambodian Pronunciation.

Bulletin of the School of Oriental and African Studies, University of London 14(1):149–174.

Hermes, Dik J. 2006. Stylization of pitch contours. In *Methods in empirical prosody research*, edited by Stefan Sudhoff, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter, and Johannes Schließer. Walter de Gruyter, pages 29–62.

Hess, Wolfgang. 1983. *Pitch determination of speech signals: algorithms and devices*. Springer-Verlag.

Hillenbrand, James and Robert T. Gayvert. 1993. Vowel Classification Based on Fundamental Frequency and Formant Frequencies. *Journal of Speech, Language & Hearing Research* 36(4):694–700.

Hillenbrand, James, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97(5):3099–3111.

Hirst, Daniel and Robert Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* 15:71–85.

Hockett, Charles F. 1947. Peiping Phonology. *Journal of the American Oriental Society* 67(4):253–267.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50(3):346–363.

Huang, Jingyuan and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America* 125(6):3983–3994.

Huffman, Marie K. 1985. Measures of phonation type in Hmong. *Working Papers in Phonetics, Department of Linguistics, UCLA* 61:1–25.

Hyman, Larry M. 2010. Do tones have features? *UC Berkeley Phonology Lab Annual Report* :1–20.

Ishi, Carlos Toshinori, Ken-Ichi Sakakibara, Hiroshi Ishiguro, and Norihiro Hagita. 2008. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1):47–56.

Janda, Laura A., Tore Nesset, and R. Harald Baayen. 2010. Capturing correlational structure in Russian paradigms: a case study in logistic mixed-effects modeling. *Corpus linguistics and linguistic theory* 6(1):29–48.

Jansen, Aren. 2008. Geometric and landmark-based approaches to speech representation and recognition. Ph.D. thesis, The University of Chicago.

Jansen, Aren and Partha Niyogi. to appear. Point process models for event-based speech recognition. *Speech Communication* .

Jongman, Allard, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America* 108(3):1252–1263.

Jun, Sun-Ah. 2000. Korean ToBI, Version 3. *UCLA Working Papers in Phonetics* 99:149–173.

Kaplan, Aaron. 2008. Peak delay and tonal noniterativity. ROA 972-0508.

Keller, Frank. 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Ph.D. thesis, University of Edinburgh.

Kenstowicz, Michael and Charles Kisseberth. 1977. *Topics in phonological theory*. Academic Press, Inc.

Kenstowicz, Michael and Charles Kisseberth. 1979. *Generative phonology*. Academic Press, Inc.

Khouw, Edward and Valter Ciocca. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35(1):104–117.

Kirby, James. 2010. Dialect experience in Vietnamese tone perception. *The Journal of the Acoustical Society of America* 127(6):3749–3757.

Kirk, Pau L., Jenny Ladefoged, and Peter Ladefoged. 1993. Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In *American Indian linguistisc and ethnolography in honor of Laurence C. Thompson*, edited by A. Mattina and T. Montler. University of Montana, pages 435–450.

Klatt, Dennis H. and Laura C. Klatt. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87(2):820–857.

Kliegl, Reinhold, Michael E. J. Masson, and Eike M. Richter. 2010. A linear mixed model analysis of masked repetition priming. *Visual Cognition* 18(5):655.

Kochanski, G., E. Grabe, J. Coleman, and B. Rosner. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America* 118:1038–1054.

Kochanski, Greg, Chilin Shih, and Hongyan Jing. 2003. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication* 41:625–645.

Krishnan, Ananthanarayan, Yisheng Xu, Jackson Gandour, and Peter Cariani. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research* 25:161–168.

Kuang, Jianjing. 2011. Phonation contrast in two register contrast languages and its influence on vowel quality and tone. In *Proceedings of ICPhS XVII*.

Kuhl, Patricia K. 2004. Early language acquisition: cracking the speech code. *Nat Rev Neurosci* 5(11):831–843.

Kuhl, Patricia K, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9(2):F13–F21.

Kuhl, Patricia K., Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America* 100(15):9096–9101.

Ladefoged, Peter and D. E. Broadbent. 1957. Information Conveyed by Vowels. *The Journal of the Acoustical Society of America* 29(1):98–104.

Ladefoged, Peter and Ian Maddieson. 1996. *The sounds of the world's languages*. Blackwell Publishing.

Laniran, Yetunde Olabisi. 1992. Intonation in tone languages: the phonetic implementation of tones in Yoruba. Ph.D. thesis, Cornell University.

Levow, Gina-Anne. 2005. Context in multi-lingual tone and pitch accent recognition. In *Proceedings of INTERSPEECH 2005*. pages 1809–1812.

Levow, Gina-Anne. 2006. Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. pages 224–231.

Li, Yujia and Tan Lee. 2007. Perceptual Equivalence of Approximated Cantonese Tone Contours. In *INTERSPEECH-2007*. pages 2677–2680.

Li, Yujia, Tan Lee, and Yao Qian. 2002. Acoustical F0 analysis of continuous Cantonese speech. In *Proceedings of International Symposium on Chinese Spoken Language Processing*. pages 127–130.

Li, Yujia, Tan Lee, and Yao Qian. 2004. F0 analysis and modeling for Cantonese Text-to-Speech. In *SP-2004*. pages 467–470.

Liberman, Mark and Janet Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In *Language sound structure*. The MIT Press, pages 157–233.

Liberman, Mark, J. Michael Schultz, Soonhyun Hong, and Vincent Okeke. 1992. The phonetics of Igbo tone. In *ICSLP-1992*. pages 743–746.

Lin, Ying. 2005. Learning features and segments from waveforms: a statistical model of early phonological acquisition. Ph.D. thesis, University of California Los Angeles.

Lin, Ying and Jeff Mielke. 2008. Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics* 14(1). URL `http://repository.upenn.edu/pwpl/vol14/iss1/19`.

Liu, Huei-Mei, Patricia K. Kuhl, and Feng-Ming Tsao. 2003. An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science* 6(3):F1–F10.

Lotto, Andrew J. and Lori L. Holt. 2006. Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics* 68(2):178–183.

Ma, Joan K.-Y., Valter Ciocca, and Tara Whitehill. 2005. Contextual effect on perception of lexical tones in Cantonese. In *INTERSPEECH-2005*. pages 401–404.

Ma, Joan K-Y, Valter Ciocca, and Tara L. Whitehill. 2006. Effect of intonation on Cantonese lexical tones. *Journal of Acoustical Society of America* 120(6):3978–3987.

MacKay, David. 2003. *Information theory, pattern recognition and neural networks*. Cambridge University Press.

Maddieson, Ian. 1977. Universals of tone. In *Universals of Human Language: Volume 2 Phonology*, edited by Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik. Stanford University Press, pages 335–365.

Maddieson, Ian. 1978. The frequency of tones. *UCLA Working Papers in Phonetics* 41:43–52.

Maddieson, Ian and Peter Ladefoged. 1985. "Tense" and "lax" in four minority languages of China. *Journal of Phonetics* 13:433–454.

Marr, David. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company, New York.

Matthews, Stephen and Virginia Yip. 1994. *Cantonese: a comprehensive grammar*. Routledge, New York.

Mattock, Karen, Monika Molnar, Linda Polka, and Denis Burnham. 2008. The developmental course of lexical tone perception in the first year of life. *Cognition* 106(3):1367–1381.

Meyer, A.R. and M.J. Fischer. 1971. Economy of description by automata, grammars, and formal systems. In *12th Annual IEEE Symposium on Switching and Automata THeory*. pages 188–191.

Michaud, Alexis. 2004. Final Consonants and Glottalization: New Perspectives from Hanoi Vietnamese. *Phonetica* 61(2-3):119–146.

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford University Press.

Miller, George A. and J. C. R. Licklider. 1950. The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America* 22(2):167–173.

Minsky, Marvin and Seymour Papert. 1971. Progress report on artificial intelligence. http://web.media.mit.edu/ minsky/papers/PR1971.html.

Mok, Peggy Pik-Ki and Peggy Wai-Yi Wong. 2010a. Perception of the merging tones in Hong Kong Cantonese: preliminary data on monosyllables. In *INTERSPEECH-2010*.

Mok, Peggy Pik-Ki and Peggy Wai-Yi Wong. 2010b. Production of the merging tones in Hong Kong Cantonese: preliminary data on monosyllables. In *INTERSPEECH-2010*.

Myers, Scott. 2003. F0 timing in Kinyarwanda. *Phonetica* 60:71–97.

Narayan, Chandan R., Janet F. Werker, and Patrice Speeter Beddor. 2010. The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science* 13(3):407–420.

Nearey, Terrance M. 1989. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America* 85(5):2088–2113.

Nearey, Terrance M. 1992. Applications of generalized linear modeling to vowel data. In *ICSLP-1992*. pages 583–586.

Nearey, Terrance M. and Peter F. Assmann. 1986. Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America* 80(5):1297–1308.

Ohala, John J. and William G. Ewan. 1973. Speed of Pitch Change. *The Journal of the Acoustical Society of America* 53(1):345.

on-line hacker Jargon File, The. 2003. Sussman attains enlightenment. URL `http://www.catb.org/jargon/html/koans.html#id3141241`.

Pelli, D. G. 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision* 10:437–442.

Peng, Gang and William S.-Y. Wang. 2005. Tone recognition of continuous Cantonese speech based on support vector machines. *Speech Communication* 45(1):49–62.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101(3):B31–B41.

Peterson, Gordon E. and Harold L. Barney. 1952. Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America* 24(2):175–184.

Pham, Andrea Hoa. 2003. *Vietnamese tone: a new analysis*. Routledge, New York.

Pierrehumbert, Janet. 2003a. Probabilistic Phonology: Discrimination and Robustness. In *Probability Theory in Linguistics*, edited by Rens Bod, Jennifer Hay, and Stefanie Jannedy. The MIT Press, pages 177–228.

Pierrehumbert, Janet B. 1990. Phonological and phonetic representation. *Journal of Phonetics* :375–394.

Pierrehumbert, Janet B. 2003b. Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech* 46(2-3):115–154.

Pinheiro, José C. and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. Springer.

Poeppel, David. 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication* 41(1):245–255.

Poggio, Tomaso, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. 2004. General conditions for predictivity in learning theory. *Nature* 428(6981):419–422.

Polka, Linda, Peter W. Jusczyk, and Susan Rvachew. 1995. Methods for studying speech perception in infants and children. In *Speech perception and linguistic experience: issues in cross-language research*, edited by W. Strange. York Press, Baltimore, pages 49–89.

Polka, Linda and Janet F. Werker. 1994. Developmental Changes in Perception of Nonnative Vowel Contrasts. *Journal of Experimental Psychology: Human Perception and Performance* 20(2):421–435.

Qian, Yao, Tan Lee, and Frank K. Soong. 2007. Tone recognition in continuous Cantonese speech using supratone models. *The Journal of the Acoustical Society of America* 121(5):2936–2945.

Quené, Hugo and Huub van den Bergh. 2004. On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication* 43(1-2):103–121.

Quené, Hugo and Huub van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59(4):413–425.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. `http://www.R-project.org` ISBN 3-900051-07-0.

Ratliff, Martha. 1992. *Meaningful Tone: A Study of Tonal Morphology in Compounds, Form Classes and Expressive Phrases in White Hmong*. Center for Southeast Asian Studies, Northern Illinois University.

Rose, Phil. 1987. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication* 6(4):343–352.

Samuel, Arthur. 1996. Phoneme Restoration. *Language and Cognitive Processes* 11(6):647.

Shi, Rushen. in press. Contextual Variability and Infants' Perception of Tonal Categories. *Chinese Journal of Phonetics* .

Shih, Chilin and Greg P. Kochanski. 2000. Chinese tone modeling with Stem-ML. In *Proceedings of ICSLP*. Beijing, China, pages 67–70.

Silverman, Daniel, Barbara Blankenship, Paul Kirk, and Peter Ladefoged. 1995. Phonetic structures in Jalapa Mazatec. *Anthrolopological Linguistics* 37:70–88.

Silverman, Kim E. A. and Janet B. Pierrehumbert. 1990. The timing of prenuclear high accents in English. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kingston and Mary E. Beckman. Cambridge University Press, pages 72–114.

Stevens, Kenneth N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111(4):1872–1891.

Strange, Winifred, James J. Jenkins, and Thomas L. Johnson. 1983. Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America* 74(3):695–705.

Sundara, Megha, Linda Polka, and Fred Genesee. 2006. Language-experience facilitates discrimination of /d-/ in monolingual and bilingual acquisition of English. *Cognition* 100(2):369–388.

Sundberg, Johan. 1979. Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics* 7:71–79.

Surana, Kushan and Janet Slifka. 2006. Acoustic cues for the classification of regular and irregular phonation. In *Proceedings of INTERSPEECH 2006*. pages 693–696.

Surendran, Dinoj. 2007. Analysis and automatic recognition of tones in Mandarin Chinese. Ph.D. thesis, University of Chicago.

Surendran, Dinoj and Gina-Anne Levow. 2008. Can voice quality improve Mandarin tone recognition? In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. pages 4177–4180.

Talkin, David. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech coding and synthesis*, edited by W. B. Kleijn and K. K. Paliwal. Elsevier Science Inc., pages 495–518.

Taylor, Paul. 2000. Analysis and synthesis of intonation using the Tilt model. *The Journal of the Acoustical Society of America* 107:1697–1714.

Tees, Richard C. and Janet F. Werker. 1984. Perceptual flexibility: maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology* 38(4):579–590.

Tian, Ye, Jian-Lai Zhou, Min Chu, and E. Chang. 2004. Tone recognition with fractionized models and outlined features. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1. pages I–105–8 vol.1.

Toscano, Joseph C. and Bob McMurray. 2010. Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science* 34(3):434–464.

Valiant, Leslie. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134–1142.

Vallabha, Gautam K., James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104(33):13273–13278.

van de Weijer, Joost. 1998. Language input for word discovery. Ph.D. thesis, Katholieke Universiteit Nijmegen, Nijmegen, The Netherlands.

van de Weijer, Joost. 2002. How much does an infant hear in a day? In *Proceedings of the GALA2001 Conference on Language Acquisition*. pages 279–282.

Vance, Timothy J. 1976. An experimental investigation of tone and intonation in Cantonese. *Phonetica* 33:368–392.

Vance, Timothy J. 1977. Tonal distinctions in Cantonese. *Phonetica* 34:93–107.

Vapnik, V. N. and A. Ya. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of probability and its applications* 16(2):264–280.

Vapnik, Vladimir N. 1995. *The nature of statistical learning*. Springer.

Venables, W. N. and B. D. Ripley. 2002. *Modern applied statistics with S*. Springer, fourth edition.

Vishnubhotla, Srikanth and Carol Y. Espy-Wilson. 2007. Detection of irregular phonation in speech. In *ICPhS-2007*. pages 2053–2056.

Wahba, Grace. 2002. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences of the United States of America* 99(26):16524 –16530.

Wang, Miaomiao, Miaomiao Wen, Keikichi Hirose, and Nobuaki Minematsu. 2010. Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model. In *Proceedings of INTERSPEECH 2010*.

Wang, Siwei and Gina-Anne Levow. 2008. Mandarin Chinese tone nucleus detection with landmarks. In *Proceedings of Interspeech 2008*. pages 1101–1104.

Wang, William S-Y. 1967. Phonological Features of Tone. *International Journal of American Linguistics* 33(2):93–105.

Warren, Richard M. 1970. Perceptual Restoration of Missing Speech Sounds. *Science* 167(3917):392–393.

Werker, Janet F. 1994. Cross-language speech perception: development change does not involve loss. In *The development of speech perception*, edited by Judith C. Goodman and Howard C. Nusbaum. MIT Press, pages 93–120.

Werker, Janet F., Rushen Shi, Renee Desjardins, Judith E. Pegg, and Linda Polka. 1998. Three methods for testing infant speech perception. In *Perceptual development: visual, auditory, and speech perception in infancy*, edited by A. M. Slater. UCL Press, pages 389–420.

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

Whalen, D. H. and Yi Xu. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49:25–47.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. Springer.

Widdows, Dominic. 2004. *Geometry and meaning*. CSLI.

Wong, Patrick C. M. and Randy L. Diehl. 2003. Perceptual Normalization for Inter- and Intratalker Variation in Cantonese Level Tones. *Journal of Speech, Language & Hearing Research* 46(2):413–421.

Wong, Wai Yi P., Marjorie K. M. Chan, and Mary E. Beckman. 2005. An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese. In *Prosodic typology*, edited by Sun-Ah Jun. Oxford University Press, pages 271–301.

Wong, Ying Wai. 2006. Contextual Tonal Variations and Pitch Targets in Cantonese. In *Proceedings of Speech Prosody 2006, Dresden*.

Xu, Nan and Denis Burnham. submitted. Tone hyperarticulation in Cantonese infant-directed speech. *Developmental Science* .

Xu, Yi. 1994. Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America* 95(4):2240–2253.

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25:61–83.

Xu, Yi and Xuejing Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America* 111(3):1399–1413.

Yang, Ruo-Xiao. 2011. The phonation factor in the categorical perception of Mandarin tones. In *Proceedings of ICPhS XVII*. pages 2204–2207.

Yang, W.-J., J.-C. Lee, Y.-C. Chang, and H.-C. Wang. 1988. Hidden Markov model for Mandarin lexical tone recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 36(7):988–992.

Yip, Moira. 2002. *Tone*. Cambridge University Press.

Yu, Alan C. L. 2007. Understanding Near Mergers: The Case of Morphological Tone in Cantonese. *Phonology* 24(01):187–214.

Yu, Alan C. L. 2009. Tonal mapping in Cantonese vocative reduplication. In *Proceedings of BLS 35*.

Yu, Kristine M. 2010. Laryngealization and features for Chinese tonal recognition. In *INTERSPEECH-2010*.

Yu, Kristine M. and Hiu Wai Lam. 2011. The role of creaky voice in Cantonese tonal perception. In *Proceedings of ICPhS XVII*.

Zee, Eric. 1998. Resonance frequency and vowel transcription in Cantonese. In *Proceedings of the 10th North American Conference of Chinese Linguistics and the 7th Annual Meeting of the International Association of Chinese Linguistics*. Graduate Students in Linguistics (GSIL) at USC, Los Angeles, pages 90–97.

Zhang, J.-S. and K. Hirose. 2000. Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 3. pages 1419–1422.

Zhang, Jin-Song, Satoshi Nakamure, and Keikichi Hirose. 2000. Discriminating Chinese Lexical Tones by Anchoring F0 Features. In *ICSLP-2000*. pages 87–90.

Zhang, Jinsong and Keikichi Hirose. 2004. Tone nucleus modeling for Chinese lexical tone recognition. *Speech Communication* 42(3-4):447–466.

Zhou, Ning, Wenle Zhang, Chao-Yang Lee, and Li Xu. 2008. Lexical Tone Recognition with an Artificial Neural Network. *Ear and hearing* 29(3):326–335.