LINGUISTIC CONTEXT AND THE SOCIAL MEANING OF VOICE QUALITY VARIATION

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Patrick R. Callier, M.S.

Washington, DC
October 21, 2013
This dissertation investigates the linguistic and social constraints on the occurrence of creaky voice quality (creak) in Beijing Mandarin (BM), as well as the effect of linguistic and prosodic context on creak’s social meanings for Mandarin listeners. It is a two-phase study, composed of 1) a production study of the distribution of creak in the naturalistic speech of speakers of BM and 2) an experimental study exploring listeners’ evaluations of creak in different linguistic environments. Not only does this dissertation expand on our knowledge of the patterning of suprasegmental variation, as well as of sociolinguistic variation in Chinese more broadly, it also opens up inquiry into the interaction between linguistic factors and the social meanings of linguistic variability.

I collected and examined a corpus of sociolinguistic interviews with 15 Beijing area residents, 18-20 years old. Using acoustic and auditory methods, I quantified the distribution of creak according to a number of factors, both social (sex, region) and linguistic (tone, position in intonational phrase). Creak is more common on low tones and toneless syllables (so-called ‘neutral tone’), as well as phrase-finally. Meanwhile, acoustic measures also indicate an interaction between tone and prosodic environment: toneless syllables in phrase-final environments are less creaky than other tones in the same prosodic positions.

I also performed a ‘matched guise’ experiment, soliciting native Mandarin listeners’ reactions to short stretches of speech with and without creak, which were
otherwise identical. These audio stimuli were designed to investigate reactions to creak in different tonal (low tone vs. high falling tone) and prosodic environments (phrase-medial vs. phrase-final). The results indicate that creak in IP-final environments indexes increased enthusiasm and interest compared to baseline. IP-medial creak, meanwhile, indexes decreased enthusiasm and interest compared to baseline. This result provides intriguing support for the hypothesis that linguistic environment plays a role in determining the social and affective meanings of linguistic variation.

INDEX WORDS: Sociophonetics, Creaky voice, Voice quality, Phonation type, Stylistic variation, Sociolinguistics, Mandarin Chinese, Beijing
DEDICATION

For Jeff
Acknowledgements

My brevity here should only be taken as a sign that I owe far too many debts of gratitude to far too many people to do them justice in a few pages they may never see. If you are not here, I know who you are—I will buy you lunch, dinner, or edit your article (in increasing order of depth of indebtedness).

Thanks first of all to Natalie Schilling, my brilliant, demanding, and compassionate mentor, for all her guidance on this sometimes bumpy project. Robert Podesva deserves credit for so much—not least hooking me into experimental work and encouraging me as I continued working on a language some audiences didn’t really want to hear about. Elizabeth Zsiga is a remarkable teacher and a perceptive, understanding advisor. Qing Zhang, meanwhile, has been a constant source of inspiration, and her scholarship is a model of the kind of work I would someday like to produce.

My friends inside and outside of the Linguistics Department have provided innumerable forms of support, love, and assistance. Anastasia Nylund stands out among these as one of the most dynamic conversation partners and dependable writing buddies a linguist could ask for. It has been a privilege to write alongside her.

My family, especially my parents Jane and Marc Callier, have held open almost every gate along the long road that has led to this point. I owe them everything, and on top of that they are actually just swell people. I hope you get a chance to meet them.

And of course, to my innumerable other advisors, teachers, colleagues, friends, and associates: thank you.
# Table of Contents

**Introduction** ......................................................... 1

**Chapter**

1  Linguistic and phonetic introduction .................................. 9
   1.1 Mandarin—structure and social questions .......................... 9
   1.2 Creaky voice .................................................. 13
   1.3 Breathy voice ............................................... 22
   1.4 Creaky voice in Mandarin: What we know ..................... 23

2  Pinning down “social meaning” ....................................... 24
   2.1 Style ......................................................... 24
   2.2 Social meaning: semiotic preliminaries ........................ 27
   2.3 A structuralist account of social meaning? .................... 29
   2.4 Stance and indirect indexicality ................................ 35
   2.5 Context-dependence .......................................... 44
   2.6 The social meaning of creak .................................. 54
   2.7 Creak in stylistic contexts and co-texts ...................... 59
   2.8 Prospectus ..................................................... 60

3  Linguistic and social constraints on voice quality ................. 61
   3.1 Participants, data and methods .................................. 61
   3.2 Results ....................................................... 83
   3.3 Discussion .................................................... 88
   3.4 Conclusion ..................................................... 92

4  Perception and the acoustics of creaky voice ....................... 93
   4.1 Background ................................................... 93
   4.2 Methods ...................................................... 96
   4.3 Results ...................................................... 100
   4.4 Discussion and Conclusion ..................................... 113

5  Linguistic context and the social evaluation of creaky voice .... 116
   5.1 Background ................................................... 116
   5.2 Methods ...................................................... 118
   5.3 Results ...................................................... 125
   5.4 Discussion and Conclusion ..................................... 128
6 Conclusion ........................................................................................................ 139
  6.1 Linguistic questions ...................................................................................... 139
  6.2 The social meanings of creaky voice ............................................................. 140
  6.3 Methodological issues ................................................................................... 143
  6.4 What we still want to know ............................................................................. 147

APPENDIX

A Pinyin-to-ARPABET mapping for P2FA .............................................................. 149
B Matched guise experiment participants and materials ........................................ 151

Bibliography ......................................................................................................... 153
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Glottal diagram showing articulatory settings for creak. Adapted from Laver (1980, 109, 123)</td>
<td>15</td>
</tr>
<tr>
<td>1.2</td>
<td>FFT spectra of my voice, pronouncing [a] in modal (solid) and creaky (dotted) phonation. Trend lines show the difference in slope between the two versions. Creaky voice has a more positive (flatter, as in almost all cases) slope.</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>One semiotic representation of the social meaning of raised /ay/ in mid-20th century Martha’s Vineyard.</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>The structuralist cookie jar fills up quickly: multiple meanings of /r/-lessness</td>
<td>31</td>
</tr>
<tr>
<td>2.3</td>
<td>Indexical field for released /t/ (reproduced from Eckert 2008, 469: “Boxes = social types, black = permanent qualities, gray = stances.”)</td>
<td>46</td>
</tr>
<tr>
<td>2.4</td>
<td>Indexical field for tag questions at a high school in the north of England (reproduced from Moore &amp; Podesva 2009, 479)</td>
<td>49</td>
</tr>
<tr>
<td>2.5</td>
<td>Indexical field for creaky voice (reproduced from Podesva 2010)</td>
<td>58</td>
</tr>
<tr>
<td>3.1</td>
<td>From Beijing Tourism Group n.d.</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td>FFT spectrum of Peter Ladefoged pronouncing schwa (<a href="http://www.phonetics.ucla.edu/course/chapter1/vowels.html">http://www.phonetics.ucla.edu/course/chapter1/vowels.html</a>). The harmonics whose amplitudes provide values for H1, H2 and other relevant measures are marked.</td>
<td>74</td>
</tr>
<tr>
<td>3.3</td>
<td>Hierarchical structure of corpus. σ denotes the syllables composing the IP; c1-5 are the 5 equidistant points at which measurements were taken in each vocalic segment.</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>Sample means for spectral tilt measurements taken at 20% intervals of vowel segments, speaker Haiyu, with increasing sample sizes. The light shaded band shows the region in which 95% of the sample means fell at a given sample size.</td>
<td>78</td>
</tr>
<tr>
<td>3.5</td>
<td>Sample standard deviations for sample means of spectral tilt measures on each speaker as sample size increases (all speakers). Note that different speakers have different amounts of data available.</td>
<td>79</td>
</tr>
<tr>
<td>3.6</td>
<td>Violin plot of tone by utterance position (width of the ‘violin’ indicates the density of data points at that value). In accordance with the nature of destressing in Mandarin (which is usually domain-final), neutral tone segments are unlikely to appear in initial position.</td>
<td>84</td>
</tr>
</tbody>
</table>
3.7 F0 plotted against spectral tilt measures, with smoothed regression lines.

4.1 Individual speakers arrayed on the vertical axis by % syllables rated as creaky. Italic face indicates male speakers, bold face female speakers. Positions are “jittered” (moved randomly) on horizontal axis for better visibility.

4.2 Standardized speaker mean H1-H2 plotted (y-axis reversed) against percent “Creaky” ratings; $R^2=0.075, 0.111, 0.091$, from top to bottom

4.3 Standardized speaker mean H1-A1 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.066, 0.09, 0.09$, from top to bottom

4.4 Standardized speaker mean H1-A2 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.007, 0.009, 0$, from top to bottom

4.5 Standardized speaker mean H1-A3 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.004, 0.006, 0.012$, from top to bottom

5.1 Screen capture of one of the questions from the online survey interface. See table 5.1 for translations of the questions.

5.2 Predicted values for Interested - Not Interested (fixed effects only)

5.3 Predicted values for Enthusiastic - Unenthusiastic (fixed effects only)
List of Tables

1.1 Tones of Beijing Mandarin ........................................... 11
1.2 Kwak’ala phonation contrasts (Gordon 2001, 387). Kwak’ala is a Wakashan language of British Columbia. ......................... 21
1.3 3-way modal, breathy, creaky contrast in Jalapa Mazatec (an Otomanguean language of Oaxaca, Mexico) (Gordon 2001, 389) ........... 21
3.1 Participants with demographic information and number of vocalic segment tokens used for analysis ........................................ 64
3.2 Five-phoneme analysis of Beijing Mandarin, with allophones (allophones that appear only as nuclei of diphthongs are marked with an asterisk) ................................................................. 69
3.3 Percent correct alignments in spot check of each speaker (approximately 100 segments per speaker) ........................................ 72
3.4 Percent correct alignments by type of segment ......................... 73
3.5 Estimates and errors from regression model of linguistic and social factors on correct alignment. The intercept represents the expectation of the log-odds of correct alignment for obstruents coming from male rural speakers ................................................................. 73
3.6 Number of vocalic segments by tonal category ....................... 83
3.7 Highest posterior density interval (HDI; 95% with bounds in columns ‘lower’ and ‘upper’) and restricted maximum likelihood parameter estimates for final model of H1-H2 ......... 85
3.8 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A1 .................................................. 86
3.9 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A2 .................................................. 87
3.10 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A3 .............................................. 88
3.11 Signs of various credible effects ....................................... 89
4.1 Creaky/Other ratings per speaker ....................................... 100
4.2 Creaky/Other ratings by tone ............................................ 101
4.3 Creaky/Other ratings by IP position (‘IP-Final’=syllable at greater than 90% of IP) ......................................................... 101
4.4 Creaky/Other ratings by IP position ................................... 101
4.5 Confusion matrix, original coding (by row) vs. recoding (by column; % ‘error’ is rate at which recode does not match original code) .... 103
4.6 Spectral tilt measures for creaky/non-creaky label categories, at each third of the rhyme. Means, with standard deviations in parentheses. 104
4.7 Confusion matrix, Linear Discriminant Analysis (spectral tilt measures only) 105
4.8 Confusion matrix, Linear Discriminant Analysis (spectral tilt measures, F0, and F1-3) 105
4.9 Confusion matrix, random forest classification (spectral tilt measures only), 500 trees, out-of-bag error estimate: 15%. 105
4.10 Confusion matrix, random forest classification (spectral tilt measures, F0, F1-3, 500 trees, out-of-bag error estimate: 14%. 106
4.11 Effect estimates of logistic regression of voice quality rating on social and linguistic factors; negative estimates favor creaky outcomes. Also given are the standard errors of the parameter estimates; the resulting z- and p-values are known to be anti-conservative, so exercise caution. 107
5.1 Response items (both poles) 123
5.2 Means and standard deviations of response values for each guise of each stimulus group. Labels are given in the same left-right order that participants saw them in. Lower values indicate responses tending toward the left-hand label, higher values the right-hand label. 125
5.3 Mixed effects model of response values for 'Charming - Not charming' 126
5.4 Mixed effects model of response values for 'High social status - Low social status' 126
5.5 Mixed effects model of response values for 'Rich - Poor' 128
5.6 Mixed effects model of response values for 'Knowledgeable - Not knowledgeable' 128
5.7 Mixed effects model of response values for 'Educated - Not educated' 130
5.8 Mixed effects model of response values for 'City - Village' 130
5.9 Mixed effects model of response values for 'Interested - Uninterested' 131
5.10 Predicted values for Interested - Not Interested (fixed effects only) 131
5.11 Mixed effects model of response values for 'Enthusiastic - Unenthusiastic' 132
5.12 Predicted values for Enthusiastic - Unenthusiastic (fixed effects only) 132
5.13 Mixed effects model of all response values 138
A.1 Mapping from pinyin to pseudo-ARPABET for P2FA, pinyin initials only 149
A.2 Mapping from pinyin to pseudo-ARPABET for P2FA, pinyin finals only 150
B.1 Participant information 151
B.2 Stimulus information 152
**INTRODUCTION**

The voice so sweet, the words so fair,
As some soft chime had stroked the air;
And though the sound had parted thence,
Still left an echo in the sense.
—Ben Jonson, Eupheme, IV.

*Vox nihil aliud quam ictus aer.*
The voice is nothing but beaten air.
—Seneca, Naturalinum Quaestionum, Book II. 29.

This project is about the social and stylistic meaning potentials of creaky voice in Mandarin Chinese, and the effect of context, both linguistic and stylistic, on the indexical properties of creak.

There are two parts to this project, broadly speaking: a production study using a corpus of Mandarin speech and a matched guise speaker evaluation experiment. The study of the sociolinguistic conditioning of non-phonemic creaky voice in a Mandarin speech corpus will lay the groundwork for experimental work, the goal of which is mainly to look at how hearers’ perceptions of creak are affected by linguistic context. This perceptual dimension and the production study together provide support for a processual and context-sensitive view of the emergence of social meaning in linguistic practice.
Why the voice?

The voice occupies an in-between place in the study of language and its sounds. The basic act of phonation—vibration of the vocal folds—is fundamental to making many of the sounds of the world’s languages. Yet the role of the entire “phonatory apparatus,” which includes the vocal folds along with a number of other structures, in helping produce the sheer variety of voice qualities one hears on a day-to-day basis remains rather poorly understood. Moreover, what ordinary language users mean by “the voice” is by no means strictly limited to the phonatory properties of a particular person’s speech. As a concept, it has significant overlap with “accent” (see [Campbell-Kibler, 2007] on what her participants called a gay ‘accent’ but I might have chosen to call gay ‘voice’), as well as bleeding outside of speech to music and non-linguistic communication. Voice quality, for many phoneticians (Abercrombie, 1967; Laver, 1980), has simply indicated those properties of a person’s acoustic performance and articulatory settings that remain the same over a long period of time, changing less quickly than the linguistically significant acoustic patterns and articulatory arrangements that allow speech to be interpreted as a series of phonetic segments.

In both China and “the West,” a person’s voice is the target of intense evaluative scrutiny. Sociolinguistics has a long history of investigating people’s stereotypes about the speech of others. The use of non-prestige languages in bilingual situations (Lambert, Frankle & Tucker, 1966), ethnic dialects (Baugh, 2003), regional dialects (Johnstone & Kiesling, 2008), and even just how someone pronounces the -ing in words like walking, sitting, and doing (Campbell-Kibler, 2007), all shape our evaluations of that speaker and his or her social characteristics.

Despite the size of this body of research, however, comparatively little is understood about the effect of a speaker’s habitual voice quality (or ‘phonation type’) on
listeners’ impressions of a talker. Much social scrutiny of voices is disproportionately targeted at women and female voices (Zhang 2008, also see Hall 1995). But the social evaluations that hearers arrive at based on listening to female voices can put women into a classic double bind (Bateson et al. 1956, 1963). Babel, McGuire, King & Satterwhite (2011), reporting experimental results of ratings of vocal attractiveness for male and female talkers in the United States, found that the reported attractiveness of a female voice was correlated with ratings of her gender typicality. The attractiveness of men’s voices, on the other hand, bore no relationship to how gender-typical their voices were. This suggests that the aesthetic value of a voice is more closely related to—and constrained by—overt gender norms for women than it is for men.

This dissertation looks at a particular kind of voice quality, so-called “creaky voice,” which has a low-sounding quality like a heavy door being opened on a ungreased hinge, and has surprisingly gotten some amount of media attention recently. Most of this recent commentary by non-linguists on “creaky voice” focuses on women, and it tends to be disapproving:

Lately, I’ve noticed a lot of young women speaking in a strangled voice that seems to be produced by a laborious effort to bypass the larynx altogether. They sound as if they are damaging their throats. […] It sounds a little Winona Ryder to me […] Also, is there some way to get them to stop? It is worse than Valley Girl intonation. (Althouse 2007)

In contrast to negative evaluations such as this, participants in a study of U.S. English described a female talker’s creaky voice as “educated,” and in a free-response task offered up adjectives such as “urban,” “young,” and “professional” to describe her (Yuasa 2010, 331-332). Yuasa hypothesizes the prevalence of creak among female speakers in her production study, combined with perceptions of them as educated
and professional, indicates creaky voice is a strategy women can adopt to sound more authoritative, which can be of value, for example, in school or at work. At the same time, negative evaluations of creak, as evidenced in broadly circulating discourses about its use among female talkers, are a liability for women who adopt this strategy.

Questions of the voice’s role in social inequality are hardly limited to gender. Representations of marginalized others, in English-speaking contexts and elsewhere, often involve the adoption of distinctive non-modal voice qualities. The harsh voice quality and extremely corpulent figure of a TV drama antagonist in a TV show which I have studied (Callier 2012), are both implicated in his portrayal as a lazy, boorish villain, and the association of his fatness with his unpleasant voice is not, I suspect, accidental. The voice is an essential dimension of the mass-mediated portrayal and disciplining of many “others”: racialized or ethnicized actors, women, people with non-normative body types, people from “elsewhere” (see the title of Barbera & Barth 2007 on voice quality in a regional variety of British English), and so on. For those of us interested in the intersections between speech, representation, and society, the voice couldn’t be a more perfect place to start.

Using Mandarin Chinese as a linguistic context (and Beijing as a broad social context), this dissertation addresses some preliminaries in the study of creaky voice and its social life. It first takes on questions around the nature of creaky voice in Mandarin, and in particular its linguistic and social distributions. Then, using experimental methodology, it delves into the question of what creaky voice’s social meanings are and how they might be affected by linguistic context.
WHY CHINESE?

East Asian contexts are themselves no stranger to evaluative and disciplinary scrutiny of the voice. Inoue’s (2006) account of discourses surrounding “schoolgirl speech” in early modern Japan provides one example. In the early twentieth century, elite male observers, hearing the linguistic innovations of an emerging class of young educated women in Tokyo, roundly rejected them as viscerally disgusting, though some of these forms were the precursor of what is widely accepted today as “Japanese women’s language.” In Japan, evaluation of women’s speech production has been consequential for the construction of the gender order itself.

Parallels to the early modern Japanese situation can be found in contemporary China. Although economic opportunity has increased in general during the rapid expansion of China’s market economy, the abandonment of socialist state feminism has increased focus on women’s physical desirability in the consumer economy and on essential difference between men and women in all areas of life (Yang 1999). The increasingly common practice of assigning women to the publicly visible “front” of a corporate setting, occupying visible positions in advertisement, reception, and secretarial work, directs attention to their outward appearance and their linguistic behavior (Zhang 2007a).

Linguistically, Mandarin Chinese offers a distinctive advantage when we go to study voice quality, creaky voice quality in particular. That is that the tonal structure of the language—the ability of pitch to distinguish word meanings—also occasions interesting patterns of voice quality variability. Tone 3, which in terms of pitch has a “low, dipping” quality is often described as creaky. As we will see in chapter 1, Chinese’s phrase-level intonational system is also of interest as regards creak.
How to study the voice, and what I found

This dissertation is a work of variationist sociolinguistics at its core. The variationist project entails uncovering the linguistic, social, and stylistic sources of variability in language and accounting for them in principled, empirically valid ways.

The first set of questions to answer is about linguistic and social distributions. In order to gather data on who uses creaky voice, and when, I conducted semi-structured “sociolinguistic interviews” with around 15 Beijing residents and recorded them in high-quality audio. Using methods I describe in chapter 3 and chapter 4, I analyzed these recordings, discovering that creak is most common at the ends of intonational phrases, on tone 3 (as well as so-called ‘neutral tone,’ which will be explained in chapter 3), and possibly in the speech of urbanite men.

The second set of questions is about what it means to use creaky voice. In order to tackle the question of creaky voice’s social meaning, I turned to a particular experimental paradigm (reported in chapter 5) in which listeners respond to creaky and non-creaky versions of the same voice (with appropriate controls to keep them from realizing that they are being tested on their evaluation of a particular feature). I find that listeners’ responses to creaky voice mostly involve perceptions of enthusiasm or interest on the part of the speaker, but that linguistic context can completely reverse these responses.

In response to these findings, especially from the experimental data, I propose in chapter 6 that the reversed interpretations of creak can both be illuminated in light of different iconic “grounds”—schemata for imbuing creak with meaning. The different prosodic contexts in which the interpretations arise predispose interpreters to one such ground or another, because of the interaction between the properties of the interpretive schema and the listener’s expectations as to the shape of the speech
stream. The choice of one ground versus the other then predisposes the listener to one evaluation rather than another.

**Roadmap and reader guide**

To describe in detail the background of these questions and the results of my investigation, the rest of the dissertation has the following structure:

**Chapter 1** contains a linguistic introduction to issues around creaky voice quality and relevant details about the linguistics of Mandarin Chinese as well as its social context. **Chapter 2** provides background to some of the relevant issues in studying the social meaning of variability in language.

**Chapter 3**, the first of two examining the sociolinguistic distribution of creak in the corpus of Beijing Mandarin I collected, conducts an analysis using measures of so-called “spectral tilt,” a well-known acoustic surrogate for voice quality. The next chapter, **chapter 4**, continues the sociolinguistic examination of creaky voice’s distribution, using hand-coded auditory labeling to identify stretches of creaky voice rather than acoustic measurements. This chapter also explores some issues in the study of voice quality in individual speakers.

**Chapter 5** delves into listeners’ subjective reactions to creaky voice, reporting the methods and results of a “matched guise” experiment aiming to assess the speaker-relative information conveyed by creak in different linguistic contexts. The results of this chapter also bring us back to questions of social meaning and its representation in sociolinguistic theory, considerations which are also addressed, alongside other incidental notes, in **chapter 6**, the conclusion.

I am writing this dissertation as a sociophonetician trained in sociolinguistic variation, heavily influenced by linguistic anthropology and interested in the processes
by which interactional meanings arise in discourse contexts. As the reader, you may be coming from a number of different backgrounds that might affect how best to approach this dissertation.

As a phonetician or empirically minded phonologist, you will note that chapter 3 and chapter 4 provide interesting data on the behavior of phonation type in a mid-to large-sized naturalistic corpus, as well as helping to relate acoustic and auditory measures of phonation type to each other. Chapter 5 also owes much to a long tradition (cf. Laver 1968) of phonetic interest in the non-referential meanings of speech sounds, especially those in the “paralinguistic” realm.

As a variationist, there are multiple points of entry to this dissertation. You will certainly see in-depth quantitative modeling of linguistic and social factors contributing to variability in speech, both from the standpoint of production and perception. In terms of theory, the perennial questions of how to delineate sociolinguistic variables, how to account for their occurrence, and how to measure their presence are all raised herein. Like many sociolinguistic variables, voice quality in Mandarin stands on a knife’s edge between participation in linguistic structure and social and stylistic variability. This work muses on many implications of this liminal status, both in its assessment of of creaky voice in production as well as in its consideration of linguistic context in creak’s perception.

If you are reading this primarily as a linguistic anthropologist, this dissertation sets out an argument about how indexical meaning arises in linguistic context, particularly where that meaning has an iconic ground, (or multiple iconic grounds, as turns out to be the case). Linguistic structure, no less than ideology or discourse context, provides paths through the thick field of meaning potentials a form has—paths we should begin to trace in as many contexts as possible.
Chapter 1

Linguistic and phonetic introduction

1.1 Mandarin—structure and social questions

Standard Chinese, putonghua, guoyu and Mandarin are all terms that can be used to refer to various standardized forms of the Beijing Mandarin dialect, instituted as a national standard in the People’s Republic of China, Taiwan and Singapore. Chen (1999) provides a good overview of the history of the formulation and promulgation of the official language standard in mainland China. Modernizers in China of the late 19th- and early 20th centuries were concerned about what they saw as China’s internal decadence and weak international position. Language standardization emerged as a modernist concern from (male) scholarly nationalist movements that had observed the “success” of standardization in Japan, then an icon of modernity in East Asia. Early standardization and linguistic modernization efforts were quite varied—the prominent modern novelist Lu Xun was known for his advocacy of Esperanto. A 1913 scholarly conference created, with official imprimatur and the involvement of linguists including Y.R. Chao, an artificial standard language that hybridized features of several large Chinese dialect groups. Eventually, however, these options were abandoned in favor of a standard based on the syntax and vocabulary of the contemporary written baihua ‘vernacular’—which was closer in grammar and vocabulary to contemporary spoken Mandarin—and on the phonology of the Mandarin variety spoken in Beijing, the Republic’s erstwhile capital and longtime cultural center of China.
1.1.1 The social life of Beijing Mandarin

Beijing itself has long been a linguistically diverse metropolis. The shape of its dialect today is understood to have taken shape over nearly a thousand years of language change, population movement and language contact (Chao 1965). Extensive work on stylistic variation in Mandarin in Beijing has been pursued by Qing Zhang (2007b, 2008, 2005, 2007a, 2006). Her work has focused on the enregisterment of cultural values associated with recognized language varieties, including but not limited to Beijing dialect, in the contemporary Chinese scene.

The local variety has several features that have strongly enregistered social meanings and are (especially together) stereotypically indexical of the Beijing variety: syllable-final rhotacization, syllable-initial lenition, and interdental realization of dental affricates (Zhang 2005). Beijing Mandarin is, meanwhile, the basis for the official standard, but these features themselves are not equally “standard” in effect. Zhang (2005) documents how rhotacization and lenition index the persona of thejingyouzi ‘Beijing smooth operator,’ a slick “Beijing native who is smooth and streetwise” (Zhang 2005, 441). Interdentalization indexes the persona of the hutong chuanzi ‘alley saunterer,’ a figure linked intimately with the distinctive alleyways of “old Beijing.”

1.1.2 Tone and intonation

Describing the sociolinguistic context in which creaky voice occurs also necessitates a characterization of the possible linguistic environments in which it might occur. The tones and intonational structure of Beijing Mandarin (BM) are of particular interest in this dissertation. Beijing Mandarin, like many if not most of the world’s languages, is a tonal language—meaning pitch dynamics can be the basis of lexical meaning contrasts. Furthermore, like many other East Asian languages, BM boasts both level
Tone Description Chao numbers Example Hanzi English meaning

<table>
<thead>
<tr>
<th>Tone</th>
<th>Description</th>
<th>Chao numbers</th>
<th>Example</th>
<th>Hanzi</th>
<th>English meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>tone 1</td>
<td>high level</td>
<td>55</td>
<td>mā</td>
<td>媽</td>
<td>‘mother’</td>
</tr>
<tr>
<td>tone 2</td>
<td>high rising</td>
<td>35</td>
<td>má</td>
<td>麻</td>
<td>‘hemp,numb’</td>
</tr>
<tr>
<td>tone 3</td>
<td>low dipping</td>
<td>21(4)</td>
<td>mǎ</td>
<td>马</td>
<td>‘horse’</td>
</tr>
<tr>
<td>tone 4</td>
<td>high falling</td>
<td>53</td>
<td>mà</td>
<td>骂</td>
<td>‘curse,scold’</td>
</tr>
<tr>
<td>tone 5</td>
<td>“neutral”</td>
<td>2 /t₁₁</td>
<td>ma</td>
<td>吗</td>
<td>[grammatical particle]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 /t₂₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 /t₃₃</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 /t₄₄</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: The tones of Beijing Mandarin, including neutral tone, along with minimal lexical contrasts between four monosyllabic words pronounced [ma]. Chao numbers for neutral tone from [Chao 1965].

Tones, which specify a level, unchanging pitch target, and contour tones, where the pitch changes over the course of the tone. There are four contrastive tonal categories in BM, as well as a fifth category, “neutral tone,” whose distribution is more limited than that of the other four and is not typically considered a full-fledged phonemic tone.

Y.R. Chao innovated the use of numbers to phonetically describe the tones of East Asian languages. Chao’s tone numerals (or tone digits) range from 1 at the bottom of the pitch range to 5 at the top of the pitch range. The four lexically contrastive tones of BM are described as having the values given in table 1.1. There are significant difficulties with using the tone numeral system precisely (see [Duanmu 2007] for an overview), but it quite economically describes the pitch dynamics that distinguish the tones of Beijing Mandarin.

Although tone in Mandarin is usually addressed as primarily a matter of pitch variation, phonological accounts of tone may offer some grist for our hypothesis-generating mill as regards voice quality. [Yip 1980] proposes two binary features to
describe tone: register, with values [+ Upper] and [– Upper], and pitch, with values [+ High] and [– High]. Register here refers to the overall “pitch register” that a tone occurs in (a coarser unit of pitch variation), while pitch refers to finer-grained pitch dynamics. One “tone-bearing unit” (TBU, i.e. a syllable, syllable rhyme, or nucleus, depending on the specific theory) in Yip’s system can have only one autosegment carrying register, but can have multiple autosegments bearing pitch. So, tone 1 (high level) units are [+ Upper, + High], but tone 2 (high rising) units are [+ Upper] register and [– High] [+ High] pitch, in sequence, to account for the rising pitch contour of tone 2.

Duanmu’s (2007) account of tone in standard Chinese (the phonology of which is based on Beijing Mandarin) mostly recapitulates Yip’s proposal, but Duanmu stipulates that register specifies phonation (in his terminology, whether a tone is “murmured” or not), while pitch specifies glottal tension (a strong articulatory correlate of F0). One important feature that unites most autosegmental accounts of BM tone is that both tone 3 and neutral tone have [– Upper] register, which corresponds to low pitch in Yip’s model and to “murmur” in Duanmu’s. Yip (1980) further stipulates that the pitch feature of neutral tone syllables is specified by spreading from previous syllables, which accounts for its different realizations after different tones.

The status of neutral tone is somewhat different in a proposal by Chen and Xu (2006), who work with a research approach propounded by Shih (1987, 2008). Shih’s approach to tone is to model both the pitch targets specified for a tone and constraints on their full implementation or partial approximation, constraints which may be specified by factors in production (e.g., speech rate), prosody (such as stress), or, possibly at the tonal level. The idea is that in certain circumstances, for instance when pitch targets specified to either side of a tone require too drastic a pitch change in too short a time, the pitch targets will be approximated rather than fully imple-
mented. When Chen and Xu apply this “target approximation” approach to the case of neutral tone, they conclude that—rather than receiving phonological pitch targets from preceding syllables or other sources, neutral tones specify their own pitch target in the middle of a speaker’s vocal range, but additionally specify a low strength of approximation to it, so that the change in pitch will typically reach this target only slowly or incompletely.

Phrase-level intonation is an area of active research in Mandarin. [Chao (1965)](1965) proposed that phrase-final neutral tones, which he assumed were unspecified for pitch, might allow intonation contours associated with different pragmatic interpretations to surface. For instance, a final neutral tone might have a rising contour in a yes-no question, but a falling one in a declarative. More recent work on differences in intonation according to speech act type ([Ho 1977](1977) [Shih 1988](1988) [Shen 1990](1990) [Yuan 2004](2004)) have found that such distinctions are reflected more in global differences in the shape of the entire F0 contour (along with other prosodic effects, such as durational adjustment) than just at the end of the IP. Nevertheless, F0 declination over the course of the Mandarin IP is well-documented ([Shih 2000](2000)). Some initial attempts have also been made at adopting the Tones and Break Indices (ToBI) transcription system to Mandarin ([Peng et al. 2006](2006)). The specific nature of prosodic and pragmatic adjustments to F0 contours, however, as well as how best to model such adjustments, remain elusive questions.

1.2 Creaky voice

This dissertation focuses narrowly on a phenomenon known as “creaky voice.” Creaky voice is an impressionistic label for a vocal quality also referred to as “vocal fry,” “laryngealization,” and “glottalization.” It is described as sounding like “a series of
taps, like a stick being run along a railing” (Catford 1964: 32, in Henton & Bladon 1988: 4), or the sound of a heavy door on an ungreased hinge.

In referring to creaky voice (or just creak) as a “voice quality,” I am attempting to be as theoretically neutral as possible and also center the perceptual aspects of the variable. The literature has sometimes preferred to use the term “phonation type” (for example, Podesva 2007 on falsetto phonation) to refer to voice qualities whose character results primarily from the state of the vocal folds; such voice qualities are distinctive because of the acoustics of the phonatory “source” wave, rather than because of “filtering” by the settings of the articulators above it. Generally most articulatory accounts (see subsection 1.2.1) posit that creak is indeed a phonatory setting involving a certain state of the vocal folds (and possibly engagement of other laryngeal structures). Nevertheless, in its social and stylistic ecosystem, creak shares the stage not just with other phonation types but with other dimensions of phonetic variability that affect what someone “sounds like”: pitch, vowel quality, loudness, etc. In their listener-centric definition, Kreiman et al. (2005) write, “voice quality may best be thought of as an interaction between a listener and a signal, such that the listener takes advantage of whatever acoustic information is available to achieve a particular perceptual goal” (345). Though this definition is extremely broad, perhaps even coterminous with the domain of speech perception, I cautiously adopt it here with a couple of limiting provisos. In the case of creaky voice, I work under the assumption that most of the relevant acoustic information is phonatory in origin (see subsection 1.2.1 and subsection 1.2.2) and that listeners’ perceptual goals may range from the strictly “linguistic” task of distinguishing different lexical tones to the “paralinguistic” aim of assessing speakers’ mental or emotional state, or even their social background.
1.2.1 The articulatory picture

There is some debate over how exactly creaky voice quality is produced, and recent work shows that there may be more than one family of settings that leads to a creaky percept. Articulatorily, creak is often described as a phonatory setting where the arytenoid cartilages are tightly adducted and there is low longitudinal tension along the vocal folds (Laver 1980, 123, see figure 1.1).

Gordon & Ladefoged (2001) arrange many of the possible voice qualities of the world on a continuum of glottal states ranging between closed (full glottal stop) and open states. Creaky voice in their model is near the “closed” end of the continuum, while modal voice (the most “ordinary” and cross-linguistically common phonatory
setting) is in the middle, with breathy voice near the “open” end. Gordon and Ladefoged themselves admit that this model is “somewhat of an oversimplification” (2001, 384), and Esling & Edmondson (2011) submit that a much more complicated assessment of the laryngeal anatomy must be taken into account to explain voice quality variation. Some instances of creaky voice, in particular, involves not only low longitudinal tension but also engagement of the aryepiglottic folds—structures that slope down and back from the epiglottis to form one of the first entry points into the laryngeal tube. In order to be distinct from so-called “harsh voice,” moreover, creaky voice must also specify the disengagement of the ventricular folds, which add noise and a tense percept to harsh voice.

Complications to the articulatory picture are also introduced by the data provided by Slifka (2006). In Slifka’s work, utterance-final nonmodal phonation was sometimes characterized by increased glottal area, indicating abduction of the arytenoids, instead of adduction as in the above models. Some acoustic hallmarks of this kind of creak are almost sinusoidal sound waves and “rapid and irregular” decay of vocal fold vibration during the closing phase of a cycle, compared to a relatively slower falloff of the acoustic wave during creak associated with vocal fold adduction.

1.2.2 The acoustics of creak

A number of acoustic phenomena have been identified as contributing to a creaky percept, including: aperiodicity, diplophonia, low amplitude between pulses of the glottis, and so-called “glottal squeak” (Redi & Shattuck-Hufnagel 2001). Aperiodicity, or an irregular spacing between glottal pulses, is one reason that creaky voice interferes with autocorrelation pitch analysis. Diplophonia is a label for alternation between higher and lower amplitude pitch periods. Glottal squeak, which usually co-occurs with at least one of the other phenomena described, is a brief period of low-amplitude,
high-frequency periodic phonation that interrupts the customary pattern of pitch periods in the waveform.

Redi and Shattuck-Hufnagel (2001) and Podesva (2010) both identify creaky voice in their production materials using a combination of perceptual and acoustic factors. By first listening to their recordings and auditorily identifying stretches where a creaky percept arose, they narrowed down the range of sites where it was necessary to apply acoustic criteria. Subsequent acoustic analysis is basically confirmatory, but is also used to measure the duration of creaky phonation. However, leaving the identification of creak’s acoustic correlates—aperiodicity, diplophonia, glottal squeak and between-pulse amplitude damping—to the ear of the researcher leaves significant room for discretion and individual variation. The relevant acoustic measures (such as jitter, for aperiodicity) are generally imprecise and have yet to supplant protocols based on inspection of individual waveforms and power spectra.

A class of measures that assess so-called “spectral tilt,” or the amplitude difference across a range of harmonics in a power spectrum (Gordon & Ladefoged 2001; Ladefoged 2003; DiCanio 2009) have played a large role in the acoustic investigation of phonation type and voice quality. In creaky voice, the high adductive tension across the vocal folds causes them to close quickly after each glottal pulse, which has the acoustic consequence of increasing energy in the upper harmonics (DiCanio 2009, 167). Various measures of “spectral tilt,” including so-called H1-H2, H1-A1, H1-A2, A1-A3 (the components of which are introduced in greater depth below), and FFT slope, have been found to distinguish voice qualities in various languages. Creaky voice should generally have more positive spectral tilt (an increase in amplitudes in higher harmonics) than modal or breathy voice, resulting in lower values on all these measures.
Figure 1.2: FFT spectra of my voice, pronouncing [a] in modal (solid) and creaky (dotted) phonation. Trend lines show the difference in slope between the two versions. Creaky voice has a more positive (flatter, as in almost all cases) slope.

Teshigawara (2003), Starr & Greene (2006) and Szakay (2008) used spectral tilt measures in sociolinguistic work and found them adequate to distinguish the voice qualities they investigated. Nevertheless, factors like fundamental frequency and vowel quality have a significant impact on these measures, so it is advisable to adopt appropriate controls when using them. The specific articulatory correlates of these measures are also somewhat different. H1-H2 (the difference between the amplitudes of the first and second harmonics in an FFT spectrum) is associated with the “open quotient,” a ratio between the amount of time the glottis is open to the total duration of an opening-closing cycle. “Longer-range” measures that cover wider spans of frequencies, particularly H1-A3 (the difference between the amplitudes of the first harmonic and the harmonic closest to the third formant), are correlated with the skewness of the glottal waveform—calculated as the ratio of time between the end of closure and the
point of maximum opening to time between point of maximum opening and start of full closure (Esling & Edmondson 2011). This dissertation will make use of the measures H1-H2, H1-A1, H1-A2, and H1-A3. Caution should be exercised in deploying H1-H2, however. Even oral vowels often involve some level of nasality, and interactions between average fundamental frequency across sexes and the relatively constant position of the “nasal pole” which affects formant heights bias H1-H2 to be lower (creakier) for men and higher (breathier) for women (Simpson 2012).

1.2.3 Perception of voice quality: establishing the object of study

The study of voice quality has traditionally addressed a fundamentally perceptual object (Esling & Edmondson 2011). Qualitative descriptions of the voice found in the work of a 2nd century Roman writer included words like *dulcem* ‘sweet,’ *pusillam* ‘feeble,’ *distractam* ‘cracked,’ and *raucam* ‘hoarse,’ among many others (Kreiman et al. 2005), which are strikingly similar to many of the qualitative labels given to voices today, even by researchers who have moved beyond purely perceptual methods (cf. Starr & Greene 2006 on ‘sweet voice’). Basing the description of voice quality on perceptual labels leads to significant difficulties. First, it is unclear how to adjudicate cases where more than one label appears to apply, and more importantly, there is no straightforward way to establish that all listeners (or which listeners) would describe the same voice quality in that way.

Laver (1980) spearheaded an effort to shift the description of voice quality to an articulatory basis. His typology of voice qualities, which is quite exhaustive and includes such labels as ‘lowered larynx voice’ and ‘protruded jaw,’ as well as ‘creaky,’ ‘harsh’ and ‘breathy’ to describe different phonation types, also included a number of ‘compound phonation types,’ which combine the elements of two or more phonation types to yield, for example, ‘whispery creak,’ ‘harsh creak,’ or even ‘whispery creaky
falsetto. A “vocal profile” approach such as Laver’s, however, is still typically based in a rater’s assessment of the sound signal itself, rather than direct observation of the relevant anatomical processes.

Kreiman and Gerratt, in a series of papers [Kreiman et al. 1992; Kreiman & Gerratt 1996 2000; Kreiman et al. 2005 2007], have established that there is wide variation across individual perceivers in their perceptions of the same vocal signals. Although the assessments of individual listeners are reasonably self-consistent—they will, for instance, largely agree with themselves when asked to rate the similarity of two vocal samples in different experimental settings—it is very difficult to achieve high levels of “inter-rater reliability” when comparing reactions to voices across raters. These problems have not discouraged linguists using perceptual methods (Henton & Bladon 1988; Belotel-Grenié & Grenié 1997; Podesva 2007 also see chapter 4 in this dissertation), and I am not saying that perceptual methods are invalid—they lie behind some of the most useful work on the sociolinguistics of voice quality to date. But we should think hard about how to describe exactly what researchers are studying when they adopt these methods, and if there is any way to enhance their validity.

1.2.4 Linguistic and social functions of creak

Creak (or creaky voice, technically the compound phonation type consisting of creak and modal voice [Laver 1980, 125]) is phonologically distinctive in many of the world’s languages, to distinguish both vowels and consonants, as exemplified in table 1.2 and table 1.3.

Creak is also used variably in many languages, such as English and Mandarin Chinese. Glottalization of word-final /t/ in English is often—though not always—accompanied by creaky voice, as are two of Mandarin’s four lexical tones:
Table 1.2: Kwakw’ala phonation contrasts (Gordon 2001, 387). Kwakw’ala is a Wakashan language of British Columbia.

<table>
<thead>
<tr>
<th>Kwakw’ala</th>
</tr>
</thead>
<tbody>
<tr>
<td>ｎａｍａ ‘nine’</td>
</tr>
<tr>
<td>ｎａｋａ ‘drinking’</td>
</tr>
<tr>
<td>ｎａｌａ ‘day’</td>
</tr>
<tr>
<td>ｎａｍ ‘one’</td>
</tr>
</tbody>
</table>

Table 1.3: 3-way modal, breathy, creaky contrast in Jalapa Mazatec (an Otomanguean language of Oaxaca, Mexico) (Gordon 2001, 389)

<table>
<thead>
<tr>
<th>Jalapa Mazatec</th>
</tr>
</thead>
<tbody>
<tr>
<td>ｊａ ‘tree’</td>
</tr>
<tr>
<td>ｊá ‘he wears’</td>
</tr>
<tr>
<td>ｊá ‘he carries’</td>
</tr>
</tbody>
</table>

the low “third tone” and high-falling “fourth tone” (Belotel-Grenié & Grenié 1997, Keating & Esposito 2006). Across many of the world’s languages, creak occurs more frequently at the ends of discourse units. Ladefoged and Johnson (2010, 150) offer that creaky voice characterizes the end of some falling intonation contours in English, which is confirmed by Henton & Bladon (1988) and Podesva (2007). In Finnish, creak co-occurs with other markers of turn completion to indicate a potential for a change of speaker (Ogden 2001).

With regard to the social distribution of creak, there is an apparent split between UK work finding more creak among men—of higher socioeconomic status in Edinburgh (Esling 1978) and lower socioeconomic status in Norwich (Trudgill 1974)—and US studies (Ingle 2005; Yuasa 2010; Podesva 2010) finding more among women.
1.3 Breathy voice

Gordon and Ladefoged’s “continuum” model of phonation places creaky voice near the “closed” end of a continuum between fully closed and fully open states of the glottis. Modal phonation lies near the middle of this continuum, and at the other pole, near the fully open (voiceless) state of the glottis lies breathy voice. Breathy voice gives an auditory impression of whisperiness, and typically contains a lot of aperiodic noise at high frequencies. It is, like creaky voice, phonemically contrastive in many of the world’s languages (for example, Jalapa Mazatec; see table 1.3). Even in languages which do not use voice quality contrastively, breathy voice is also a very common phonation at the ends of intonational phrases (Smith 2002; Ogden 2004).

Articulatorily, breathy phonation is generally understood to be produced by widely abducted (separated) arytenoid cartilages, which may or may not be accompanied by enough longitudinal tension to produce voicing (“breathy voice”; Laver 1980; Esling & Harris 2005). Acoustic measures of spectral tilt (H1-H2, H1-A1, H1-A2, H1-A3), which we above explained are inversely correlated with the presence of creaky voice, can also be used to diagnose breathy voice. Breathy voice is characterized by a higher open quotient (the glottis is open longer compared to the amount of time it is closed), which reduces energy in upper harmonics and thus results in more positive values on all these measures. Although the main object of study in this dissertation is creaky voice, breathy voice does make a cameo appearance in chapter 3 in the course of my investigation of the linguistic and social conditioning of voice quality in Beijing Mandarin.
1.4 Creaky voice in Mandarin: What we know

Phoneticians and phonologists widely agree that Mandarin tone 3 (the “low dipping” tone) is often associated with non-modal phonation. Duanmu uses the term *murmur*, which is confusing because it is normally associated with breathy voice. Belotel-Grenié & Grenié (1997) calls it *craquée* ‘cracked/creaky,’ and also notes a certain amount of the voice quality on tone 4 (high falling, cf. Xu et al. 2012 who also find creaky voice on tone 3 and tone 4). “Creaky” is also the label given to tone 3 in Tianjin Mandarin (closely related to Beijing Mandarin), where it is realized as a low rising tone (13). Tianjin tone 3 sounds “creaky,” but in fact has values of $H1 - H2$ that correspond to breathy phonation (Davison 1991). As for creak’s role in intonation and prosody, Belotel-Grenié & Grenié (2004) provide suggestive evidence that creak is used to mark the ends of intonational phrases and longer discourse units in Mandarin.

Xu Xiaoying (Xu et al. 2012) has begun a sociolinguistic investigation into the occurrence of creaky voice in a large corpus of Mandarin speech (recordings of an orally administered Mandarin proficiency test). Her findings, based on auditory coding, do not suggest a gender difference in the incidence of creaky voice among Mandarin speakers, but there is a positive correlation between participants’ Mandarin proficiency and the amount of creaky voice they use. Suggestively, there is also a marked difference in the instance of creak between history majors (30%) and all other majors, be they technical, scientific or humanistic (13%-22%). Nonetheless there is considerable room for more research into creaky voice: not just with regard to linguistic and social constraints on its distribution but also into stylistic variation and the social meaning of variability in voice quality (and suprasegmental variation more generally). Chapter 2 will introduce us to some of the theoretical issues in investigating these last questions—style and social meaning.
Style is a dimension of linguistic variation that attracts attention for many reasons—most prominent scholars of style in linguistics seem to come at it with their own distinct set of motivations. Besides puzzling out when and why speakers might speak differently from one moment to the next, or exploring the implications of intraspeaker variation for theories of grammar, many researchers have begun to probe issues stemming from the nature of style as a meaningful aggregate—that is, a collection of signs experienced together. Thus, fleshing out the new complexities of meaning in third-wave and speaker design approaches to style requires a detour through the semiotic basis of our enterprise. This will help us make explicit some of our assumptions about how so-called “social” meaning works and thereby point to new questions to be answered about how it functions.

2.1 Style

The variationist study of style originates in Labov’s early work (Labov 1972a, Ch. 4) in New York City, where in the context of the semi-structured sociolinguistic interview, he quantified rates of variable processes according to the interactive context and the genre of speech being produced. So Contexts A1-5 (speech outside the interview, with a third person, not in direct response to questions, childhood rhymes and customs, danger of death) contrast with Context B (“interview speech” or careful speech),
on down to Context C (reading), and Context D and Context D’ (word lists and minimal pairs, respectively). Labov differentiates these contexts to allow sociolinguists to asymptotically approach their speakers’ vernacular repertoires (evident in “Style A,” casual speech).

Labov discovered that the realization of many variable features shows differentiation across these contexts. For instance, he found that, across social classes in New York City, rates of postvocalic /r/-deletion decreased steadily as you moved from Context A to Context B, from B to C, and from C to D and D’. This effect held for ‘G-dropping,’ /r/-lessness, and a number of other variables.

Labov’s axiomatic understanding of style (at least in the sociolinguistic interview) was that it is organized by the level of attention paid to speech in any given context. As speakers increase their attention to the form of their utterances, this model predicts that their use of stigmatized variants will decrease and of prestige variants will increase. Style as attention to speech is, in this way, a unidimensional model. Other early theories of style, Bell’s (1984) ‘audience design’ chief among them, rejected attention-to-speech as a ‘nonstarter’ for explaining much or even most stylistic variability. Citing a lack of evidence for Labov’s psychological claims and using data from radio programs where attention paid to speech was presumably invariant, alongside other studies of language variation in relation to audience, Bell’s proposal documents how speakers’ frequencies of certain sociolinguistic variants can be observed converging with or diverging from those of their interlocutors, or to converge with an audience’s perceived ideal.

In situations where style shift is observed without obvious effect from present interlocutors or audience members, in particular in cases of initiative rather than reactive style shifting, Bell introduces the construct of ‘referee design,’ denoting the capacity to converge with the linguistic production of an imagined audience of some
significance to the speaker. Audience design introduces an interesting problematic about metalinguistic knowledge: how do talkers know which features to manipulate, and to what extent?

In the case of design for a copresent interlocutor (who provides his/her own speech ‘data’ to the talker), we sometimes assume that this knowledge can be built up over the course of interaction. This kind of audience design is the kind accounted for by online accommodation. Accommodation to nonspeaking audiences (e.g. overhearers) or ‘initiative shifts’ accommodating to imaginary referees, however, requires metalinguistic knowledge from outside the limits of discretely bounded interactions.

Recent approaches to style have focused on ways in which style shifts rely on interactants’ (usually implicit) metalinguistic awareness of the semiotic associations of individual forms. These more current approaches, put forth under the headings of “speaker design” (Schilling 2013; Coupland 2007) and “third wave variation” studies (Eckert 2012) respond to two broad areas of concern about older theories. Studies associated with the “speaker design” family of theories take on the hegemony of reaction in Bell’s design and account for the “initiative” dimension of some style shifts. The “third wave” of sociolinguistic variation studies also takes speaker agency as a starting point, and focuses on identifying and describing the full range of stylistic functions a variable can take on. A stylistic variant’s meanings are not just limited to associations with sociodemographic categories or locally salient social groups, as in the first two waves, but can be associated with affective or epistemic stancetaking, take on “iconic” or at least relatively “motivated” meanings, and be linked to social dynamics within individual interactions.
2.2 Social meaning: semiotic preliminaries

The heritage of sociolinguistics as the grandchild of large-scale regional dialectology has left its mark on linguists’ approaches to the social meaning of variability in language. Eckert (2012) has proposed a model of variation studies coming in three waves, roughly ordered in time. Early studies fall mostly into Eckert’s first and second waves of variation studies, approaching quantitative variation as (a) conditioned by macrosociological factors in the first wave (ostensibly affording a high degree of comparability across communities), or (b) ethnographically salient, “locally valid” categories in the second wave, which are faithful to community members’ own conception of their social reality. In assigning social meanings to the variables they study, first and second-wave studies assigned meanings of membership in or affiliation with the groups that used them the most, an approach somewhat disparagingly labeled “correlationist” by some later critics.

The first- and second-wave views of social meaning connect variants directly to the populations who use them, differing mostly in terms of the scale and ethnographic validity of the categories employed. As Eckert (1999) points out, though, if a female member of the burnout subculture in a Detroit-area high school—which rejects the quasi-corporate culture of the school environment and embraces smoking, fighting, and an urban, working-class orientation—uses a raised nucleus for pre-voiceless /ay/, it is not necessarily an act of identity proclaiming affiliation with groups like “woman,” “burnout,” or even “burnout girl.” Its increased height in the word *fight*, for instance, may be attributable to semiotic associations between the raised form and values associated with fighting, like toughness, which may be at issue in the speech contexts where *fight* is likely to arise. While a girl’s identity as female is hardly at stake in most interactional contexts, and her identity as a burnout is usually at least provisionally
secure, her construction of a “tough” persona may be entirely contingent on her performance in a particular speech activity. Thus when we examine the deployment of individual tokens of a stylistic variable, we search for meanings of this sort—what is this variant doing here, for this speaker?

This is the project laid out for Eckert’s third wave of variation studies—discovering how the meanings of intraspeaker and interspeaker variation are constructed in unfolding interaction. Though these continue to be referred to as “social” meanings, compared to the first two waves of variation studies, this approach widens considerably the range of values a variable can take on—from durable social or group-associational categories like “man” down to fleeting stances or interactional meanings like “combative.”

Perhaps not unexpectedly, this move has complicated our picture of what is actually going on behind the semiotic curtain when we see stylistic variation. In first- and second-wave variation studies, it usually suffices to say that a variable possesses a unitary, discrete, (synchronously) stable meaning. These meanings, which early variationist work also recognized could change over time and across social categories in a single speech community, pose a similar challenge for variationist semiotics to that which formal language change itself presented for structural linguistics (and which [Weinreich et al. 1968] sought to push past). Nevertheless, meaning was not theorized as such in early variationist work, (though see [Laver & Trudgill 1979] for one of the first engagements of variationism with extant semiotic theory—and an isolated example, at that, not linked into the ongoing study of formal linguistic change). For expository purposes we can speak tentatively of a semiotics adumbrated, though not explicitly theorized, in “early” variationist sociolinguistics. This proto-theory (which can be attributed to no actual researcher) bears some resemblance to Saussurean structuralism, though, to repeat the above, the empirical details about variation and
change in meanings within communities trouble this formulation. As I will review below, theories of stylistic and social meaning do not fit with structural semiotics, but why? What do we need from a theory of signs in order to faithfully account for stylistic meaning? A review of the basis and limitations of structuralism will begin to answer these questions.

2.3 A STRUCTURALIST ACCOUNT OF SOCIAL MEANING?

Saussure’s sign consists of two parts: a signifier and a signified. The signifier (sometimes translated as signal), is what conveys the meaning—the word tree, for example, or the English possessive morpheme ’s. The idea of the tree, or the semantic dimension of the possessive relation, is picked out from the undifferentiated conceptual morass by the signifier, and is known as the signified (sometimes translated as signification).

The structuralist “cookie” diagram showing the relationship between signifier and signified for Labov’s second-wave analysis of Martha’s Vineyard /ay/, then would look something like figure 2.1.

2.3.1 LIMITATIONS: POLYSEMY, INDETERMINACY, CHANGE

Saussure’s semiotics was highly influential, spreading outside of the discipline of linguistics (which it helped start) to be picked up in anthropology and later in literary studies and critical theory. But in the intervening years its limitations have been widely recognized. For us, it is enough to note that structuralism assigns a unitary and invariant value to each sign, fixed by the position of the signifier and signified in the system of differences and oppositions. This makes it hard to deal with polysemy, for instance. The lexeme tree sometimes means “a woody perennial plant typically with a single stem or trunk growing to a considerable height and bearing lateral branches”
and sometimes means “a diagram with a structure of branching connecting lines, representing different processes and relationships” (Soares & Stevenson 2008). A strict structuralist account will be forced either to posit two completely independent signs that incidentally happen to have a signifier of the same form, erasing the relationship between the two denotations of *tree* (the “polysemist” position François 2008), or to posit a single core meaning which requires additional, unspecified theoretical machinery to provide the correct meaning in context (a “monosemist” position).

The kinds of meanings we are dealing with in stylistic variation are of a different sort than lexical denotations, but variationist sociolinguistics abounds with cases exemplifying the one-to-many semiotics that gives structuralism such problems. Many sociolinguistic variables, for example post-vocalic /r/-lessness, are widely known to have multiple social meanings. Members of many North American speech communities perceive /r/-lessness as a stereotype of both English from England and English of the
northeastern US. Meanwhile there is no evidence that most people in North America would perceive, for instance, Londoners and New Yorkers as the same group, but /r/-lessness is easily recognized as characteristic of both groups. To theorize this divide in a structuralist semiotics requires, as in the “tree” thought example, positing two “signs,” as in figure 2.2.

Lexical ambiguity and polysemy, with the significant barriers they erect along the path to theories of lexical semantics and cognitively realistic grammars, have received and continue to receive ample scholarly attention, and lexical semantics has long since moved past purely structuralist frameworks. For variationists interested in social meaning, however, variability of the sort above has only just started receiving sustained scrutiny.

Podesva and Chun (2007) Chun & Podesva (2010) give several examples of indeterminacy and multiplicity in social meanings (and theorize it using concepts to be described below). They show how the vague meanings of some variants are resolved in linguistic, situational, or interactional contexts. Other deployments of stylistically
meaningful variation, however, rely on the indeterminacy of the resources in play and remain at least partially unresolved between two or more possible meanings. The performance of a “Mock Asian” style by comedian Margaret Cho, for example, relies on its ability to activate both “racist” and “subversive” meanings, and it is not possible to disambiguate between the two even at the level of individual speech acts.

Indeterminacy or sustained ambiguity is possible for other variables as well, like /r/-lessness. The storied “Boston Brahmin” accent, which is also characteristically /r/-less, is featured in the educational video American Tongues (Alvarez, Kolker, John & Media 2005). The video describes the Brahmin accent as “British-sounding” before cutting to a clip of two tweed-jacketed Brahmins discussing the novels of Charles Dickens, highlighting its Insular overtones at the same time as it describes the accent as a kind of “American” speech. Variable /r/-lessness is most rightly considered a different variable across Boston and Insular English speech communities, but in the documentary context, the “Englishness” of the Brahmins is intentionally played up, perhaps with help from their /r/-lessness. In another context, the use of /r/-lessness by Lou, a Lumbee Indian of Robeson County, NC, when discussing the Civil War, seems at times to borrow on a stereotype of highly /r/-less antebellum Southern English (Schilling-Estes 2004). At the same time, this feature is also characteristic of the African American English of the county, and at other points in the interview, Lou’s levels of /r/-lessness decrease, in divergence from the usage of his African American interlocutor, possibly in response to racially loaded meanings of the styles at play. When looked at in social and stylistic contexts (alongside co-occurring accent features and other discursive phenomena), we can separate many of these cases of /r/-dropping into clear instances of different enregistered voices. The relevant social meanings are attached to the registers, not to /r/-dropping itself. But the similar structural status of /r/-dropping in British Received Pronunciation, Boston Brahmin
talk, Robeson County AAVE, and stylized antebellum Southern English, as well as its overall salience, leads necessarily to questions about how the various meanings of these registers are activated (or not) in the presence of /r/-dropping.

These examples show that the associations between particular linguistic features and the meanings they activate are highly context-dependent and potentially multiplex. Ideally, sociolinguistics should adopt a theoretical view that can acknowledge and account for these characteristics. The most promising proposals in this vein move away from pure Saussurean structuralism and borrow theoretical machinery from Peircean semiotics, as found in, for instance, Silverstein (1976, 2003). Particularly important is the concept of indexicality. An index is a modality of sign-relation in C.S. Peirce’s theory of sign\(^1\) which includes signs that “point to contexts of occurrence” (Silverstein 2003, 195) or, more broadly, signs where the relationship between sign and object is characterized by “connection” (Parmentier 1994, 17). The work of Peirce, a logician and philosopher in the American Pragmatist tradition, found its way (back) into North American language scholarship through the work and influence of Roman Jakobson (see, for example Jakobson 1959 and also Liszka 1981), who re-discovered Peirce after his emigration to the United States. Peirce’s theory of the sign actually predates Saussure’s and consists of three parts: the sign, the object, and the interpretant.

For the sake of concision, the sign can be associated with Saussure’s signifier, and the object with Saussure’s signified. The interpretant, however, brings in a dimension that is absent in Saussure’s structuralism—it is, essentially, a sign user’s knowledge that a particular sign-object relationship counts as a sign. Any object can be represented by a theoretically unbounded number of signs. The concept of the interpretant

---

\(^1\)Peirce’s semiotics emerges over an extremely large and variegated set of primary texts. Peirce (1878, 1998) are accessible and reasonably representative works.
is a way of limiting the possible field of active, known signs. With this extra leeway, Peirce allows us to account for a single sign representing multiple objects. Silverstein (2003) theorizes this capacity in depth with regard to a number of sociolinguistic facts, but we can first illustrate with a non-linguistic example. The feeling of warmth on one’s face can be taken as a sign of the sun:

Warmth→Sun

The relationship of warmth to the light of the sun is understood as (indexically) significant by ordinary sign users. This relationship, however, can be expanded upon further. Feeling the warm sunlight on one’s skin can also signify that it is summer (or a nice day, or a good time for a walk, etc.). We can schematically represent this relationship as

[Warmth→Sun]→Summer

The idea that warm sunlight is associated with summer leads to the recognition of warmth—in certain sunny contexts—as a sign of summer in its own right. The extra traction this gives us allows for ambiguity, change, and elaboration in the meanings taken on by indexical signs.

This is precisely the picture painted by Silverstein (2003) in his explication of “indexical order.” Indexical order (not to be confused with the “orders of indexicality” proposed by Blommaert 2005) is a framework that shows how indexical meanings build on each other and are transformed in the “micro-realtime” of interaction. Just as the idea of warm sunlight can be transformed into a knowledge of the arrival of summer, many linguistic indexical forms can be transformed in the course of interaction, taking on new meanings through the mediation of other sign arrangements. The basic idea is this: that given any indexical meaning of a semiotic form (called its
“nth-order” construal), a higher order “$n+1$st-order\footnote{Higher orders can be added arbitrarily, leading to $n+1$st-order construals, $(n+1)+1$st-order construals, $((n+1)+1)+1$st-order construals, and so on.}” meaning can be constructed by filtering the $n$th-order meaning through the co-text, context, and/or broader ideologies. This process can be repeated multiple times even within a single interactional instant to yield effective meanings far removed from the $n$th-order meaning with which the form entered the interaction.

2.4 STANCE AND INDIRECT INDEXICALITY

In third-wave variation study, indexicality and indexical order have often been understood through the lens of Ochs’\footnote{Ochs’ work on the links between linguistic forms and identity categories, specifically gender. In Ochs’ account, most indexical resources used for signifying gender identities of interaction participants are indirect, in that they are actually used locally for the construction of stances, acts, and activities. Forms acquire gendered meanings through rubrics of valorization (ideologies, contexts, co-texts) that associate the stances in question—rather than the forms themselves—with gendered social categories.}

An example Ochs gives to describe this situation is a gendered paradigm of non-referential affective particles in Japanese. The particles $ze$ and $wa$ can be added to the ends of sentences in Japanese to signal stances of “coarse intensity” and “gentle intensity,” respectively. Though Japanese speakers understand $ze$ to signify the speaker is male, and $wa$ to signify the speaker is female, in Ochs’ analysis it is actually the qualities of coarseness and gentleness that the forms index, and these are in turn ideologically linked to femininity and masculinity. Thus $ze$ and $wa$’s evident, reportable indexicalities are actually mediated by stance and ideology.
In terms of indexical order, this account posits that stance-related meanings like “gentle intensity” or “softness” typically enter interaction as presupposable, “n-th-order” indexicalities of wa. On top of this foundation, and taking into consideration ideologies that associate soft and gentle qualities with women, an “n + 1st-order” construal of wa arises, one which pragmatically entails that the gender of the wa-user is female. Crucially, the genderedness of the form is always mediated by its association with stance—in this case, a “gentle” stance.

Kiesling (2005; 2009) has suggested that the default, lower-order social meanings of sociolinguistic variation are characteristically stance-related. Meanwhile, Eckert (2008), in addition to widely citing Ochs’ stance-centric analysis of language and gender, has suggested that indexical meanings of association with stable sociodemographic categories are usually higher-order indexical construals of other, local meanings.

Inoue (2006) claims that language ideologies around Japanese women’s language (JWL), including wa, rely on a putative “history effect” (76) that posits, rather than demonstrates, the stance-centric teleology Ochs describes. In the case of Japanese women’s language, there is no historical reason to attribute wa’s present-day indexicality of “softness” to its original gender association, says Inoue, because at the time, the features of JWL were heard as brash and vulgar. By positing that it was speaking women’s stance-taking agency, rather than the socio-historical emergence of the modern Japanese woman, that invested JWL forms with the meanings it has today, ideology achieves an “indexical inversion” in this case. This inversion mystifies historical links between forms like wa and feminine identity by inventing a history of its own, and makes them seem more natural and less arbitrary than they historically were.
This is a radical revision of Ochs’ claims, but by itself we might think that such “inversion” is just something ideology does every once in awhile, with stance-first indexical meanings remaining the more common case. There has not been enough attention to this particular question to say for certain, but cases of meaningful variability that do not involve stance-related nth-order meanings abound. Agha (2007a, 161) cites an example of a male Lakhota speaker’s use of wele:, a grammatical particle that also indexes that the speaker is female. In deploying this female-indexing particle, the male speaker creates an emergent effect of “motherly concern.” In another example analyzed by Coupland (2007, 156-163), Welsh politician Aneurin Bevan deftly interweaves phonetic variants associated with Received Pronunciation and Valleys Welsh English. The deployment of each variety activates associated personae and cultural values, a common pattern in dialect stylization. The male Lakhota speaker’s momentary “crossing” into women’s speech and Bevan’s juxtaposition of RP and Welsh accents put into play indexical orders that are the reverse of the one Ochs sketches for ze and wa. In both cases, the “stance” adopted by the speaker emerges as a creative n + 1st indexical effect, while the more completely sedimented social category (“woman” or “Valleys”) indexed by the form is a lower-order pragmatic presupposition.

[[wele::]→women’s language]→“maternal concern”

[monophthongal o, e; centralized au, ai; etc.→“Valleys”]→“vernacularity,” “frankness”

The stance-first approach to language and identity emerges from valid concerns, central among them an intuition about speaker strategy. We want to avoid positing that in employing a feature with the value of, say, “speaker=female,” a speaker is actively making a claim of identity in the indexed social category (a view cautioned against by, for instance, Eckert 2008). This is, perhaps, in a desire to distance ourselves
from the broad-strokes “correlational” semiotics of first- and second-wave variation studies, which by and large did not take advantage of the potential for “infinite semiosis” (i.e. indexical order) in Peircean theory. Caution is quite acceptable given the continuing background hum of theories of identity as “claim-staking” that have been implicated or explicitly adopted in sociolinguistics—theories that often possess numerous redeeming features (cf. Le Page & Tabouret-Keller 1985; Cameron & Kulick 2005) but venture too far from the facts in search of speaker agency.

Still, if we do not understand indexical meaning as necessarily tantamount to claim-staking, then we lose this motivation for a stance-centric view. Indexical meaning, especially of the sedimented or enregistered sort, is a form of pragmatic presupposition (Silverstein 1976; Ochs 1993). Forms like Lakhota wele: presuppose meanings, e.g. “speaker=female,” that can easily be counteracted, transformed, or canceled outright when we take the overall “texture” of a semiotic performance into account (Perrino 2002; Agha 2007a). Indirect indexicality is a useful corrective to “correlational/claiming” theories of language and identity, and well-evidenced examples of it abound in sociophonetics (Eckert 1999; Podesva 2006). Nevertheless, fear of correlationism does not mean we should conclude that stance always comes “before” sociodemographic categories in accounts of indexical order.3

3Some theorists have attempted to assimilate indexical order to timescales beyond the single interaction, which Silverstein does leave room for. However these efforts have often fixed particular orders (1st-, 2nd-, 3rd-, etc.) with predetermined social meanings or types of social meanings. This may have use as a general theory of change in indexical meanings across larger timescales, but it will, as the example of indexical inversion from Inoue shows, always fail to account for the details in some cases. In my view enregisterment provides a more useful and flexible vocabulary for describing the changes which indexical meanings may undergo.
2.4.1 Enregisterment

Silverstein’s account starts from the time at which a form is deployed with its \( n \)th-order meaning. What sociohistorical machinery must be in place for this to happen? Quite a bit. First a form needs to become widely recognizable as able to enact that meaning. This is not an automatic process. In a history of Japanese Women’s Language, Inoue (2006) points out that \( wa \), among other now-widespread indexes of female speaker, was associated\(^4\) with young schoolgirls in late 19th-century Tokyo, and was evaluated as crass and vulgar. Later in the 19th and early 20th-century, it came to be associated with a particular brand of Japanese femininity, and associated values of softness. This was only possible through mass-mediated discourses associating what was then known as \( jogakusei kotoba \) ‘schoolgirl speech’ with emerging portraits of the idealized urban Japanese woman. With widening circulation, this portrait of womanhood and the linguistic forms paired with it became naturalized and universalized. \( Wa \) and other forms subsequently became recognizable as “women’s language,” with a widely circulating indexical meaning of “speaker=female.”

The sociohistorical process of cementing the indexical values of a form and giving them a degree of currency across a population of language users has been called enregisterment (Agha 2005, 2007a). Enregisterment, which deals broadly with the historical organization of linguistic forms into registers, or meaningful repertoires, imbues a form with a recognizable effect, i.e. something that can act as an \( n \)th-order indexical meaning. The \( n \)th-order meanings which are enregistered for a form can be activated essentially “as-is,” as when the Queen of England speaks in a recognizable English accent and (non-British) hearers think, “Aha! She’s British!” Alternatively, enregis-

---

\(^4\)This “association” was not, Inoue (2006) argues, necessarily constructed by the actual linguistic agency of women; instead it was constructed by the privileged male “listening” subject.
tered meanings can be caught up in the vicissitudes of indexical order as described above, and in the same speech situation a hearer could activate ideological associations with British accents to arrive at a $n + 1$st-order construal of the Queen’s accent expressible as: “She sounds quite proper!” Putting enregisterment at the forefront highlights metalanguage, or talk about talk (see Lucy 1993), since metalanguage and metapragmatic discourse (talk about behavior) provide a means for knowledge about the distribution and function of linguistic forms to endure beyond individual interactions and circulate within and between speech communities.

Metalanguage and metalinguistic practices are also tied to a perennial concern of variationist sociolinguistics: that of speaker/hearer “awareness” of variation. Voice quality is particularly interesting in this respect. In different social contexts it ranges from being a primary component of highly enregistered named styles (e.g., “respectful voice” of Lachixio Zapotec [Sicoli 2010]) to being extremely difficult to talk about with non-linguists, even after extended description (as Mendoza-Denton 2011 describes in her interactions with informants). As I outline in the introduction, discourses about creaky voice in US English have begun to circulate in journalistic media and blogs under the banner of “vocal fry.” Still, in most cases when I attempt to describe my object of study to non-linguists, it requires extensive elaboration and scaffolding, including exemplification. This range of variability in how easily people talk about features is a matter of its “availability,” in Preston’s (1996) terminology, to folk linguistic awareness.

Mendoza-Denton (2011) characterizes creak in Silverstein’s (1981) related framework for describing constraints on metalinguistic awareness. Creak is lacking in “continuous segmentability” (coming all together in the speech stream, separable from what comes before and after it), “relative presuppositionality” (propositions that are necessarily presupposed by its deployment) and “unavoidable referentiality”—all
factors which predispose creak to being difficult to discuss in ordinary language. Another framework for talking about language users’ awareness is the indicator-marker-stereotype trichotomy (Labov 1972a). Indicators are sociolinguistic variables that show social differentiation but not stylistic differentiation, while markers also show stylistic differentiation but are not subject to overt comment. Stereotypes combine social and stylistic differentiation with availability for overt comment by speech community members. These categories are typically applied to features undergoing change in progress, but can also be used to talk about stable variation. As a way of assessing the metapragmatic status of variability given only the amount of interactional ethnographic context afforded by an hour-long sociolinguistic interview, this framework is impressively productive. Still, later frameworks, like Preston’s and Silverstein’s (as well as the idea of enregisterment more broadly), provide finer tools for dissecting precisely what kinds of metapragmatic activity language users are engaged in, if the researcher has other sources of information from which to triangulate. In matters of interpretation, Preston and Silverstein also avoid sticky quandaries around how to say what language users are “aware” or “conscious” of and instead try to describe how language users behave overtly with regard to the variables (though of course the criteria for Labov’s classifications are also based on overt behavior).

Nevertheless, in choosing a variable which shows a low degree of availability, which appears to be the case for creak in Mandarin as well as English, we are in a way circling around to Labov’s (1972 [1963]) dictum that the best sociolinguistic variable was one that its users were not aware of. Originally this requirement was intended to allow the study of change in progress to be addressed with interview methods that might otherwise put language users “on their guard” and cause them to “change” their speech. Today, however, in coming back around to low-awareness, low-salience, or not-highly-enregistered variables, the question is different: how can a form mean when its
users cannot name it and do not pick it out as exceptional even when prompted to do so? As the literature on enregisterment argues, however, meaningfulness precedes, perhaps necessarily so, enregisterment and awareness of a form (Agha 2005, 2007a).

2.4.2 Style as aggregate

The sense of “register” as something whose meanings are “enregistered” or recorded exists alongside another sense in which language researchers often use the term—to signify a range or repertoire of forms (a sense which underlies most sociolinguistic work on register, e.g. Biber 1988, among others). Recent style research has affirmed the axiom that stylistic meaning is fundamentally about meaning in the aggregate (Half Moon Bay Style Collective 2006). Zhang (2005) describes the features of an emergent “Cosmopolitan Mandarin,” and Podesva (for instance, 2008) demonstrates that the aggregate whole of certain styles, such as the “diva” persona of one of his speakers, are a creative product of the meanings of their component parts. This is an aspect of style that is often overlooked by studies that look at style in terms of individual variants. Podesva, Jamsu, Callier and Heitman (n.d.), for instance, conduct a sustained investigation into modulations of meaning related to variability in an exceptionally limited range of variable phonetic realizations—release/aspiration or lack of release/flapping of the phoneme /t/ in American English. These modulations are clearly meaningful, as the authors demonstrate. But a “trees-not-forest” bias is present in sociophonetics.

The study of enregisterment and the social histories of speech repertoires cum repertoires highlights how the kind of compositionality exhibited in the “diva” style is just one end of a continuum, at the other end of which lie styles or repertoires that are meaningful in and of themselves, regardless of the meanings of their constituent “features.” Voice quality is attested in both types of cases: compositionality
in its role as a “semiotic hitchhiker”—that is, an otherwise meaningless ornament to a multimodal stylistic performance (Mendoza-Denton 2011)—or as part of enregistered “respectful voice” (Sicoli 2010). Agha’s (2003, 2007a) accounts of Received Pronunciation, Inoue’s (2006) genealogy of women’s language, and Silverstein’s (1979, 2003) explication of honorific registers in Javanese, all look at registers as cultural objects which circulate as repertoires, are evaluated as repertoires, and in many respects typify the forms contained within them as the “same” (other examples in this vein include Remlinger 2009 on the “Copper Country English” of Michigan’s Keweenaw peninsula). Meanings are ascribed to these registers; they evolve, are transformed, and vanish, but—at least in these scholars’ accounts—register-mediated meanings are not particular to the individual components of the repertoire themselves.

Sociolinguistic uptake of the enregisterment concept by scholars like Zhang (2008) and Johnstone and Kiesling (2008) has looked at enregisterment of both varieties and of individual variables. Certain features of “Beijing Mandarin” (itself an enregistered variety) have circulating meanings: rhotacization invokes the persona of the jing youzi ‘Beijing smooth operator’; interdentalization of dentals the hutong chuanzi ‘alley saunterer.’ /aw/-monophthongization, a distinctive feature of Pittsburgh English, has a particular profile of meaningfulness (or lack thereof) for evaluators who have the feature in their own speech, versus those who don’t, who more uniformly hear the feature as indexical of “local” identity. These one-variable-at-a-time perspectives add a lot to our picture of socially meaningful variation, and we should never be afraid of zeroing in on individual variables. Nevertheless, we should be open to the possibility of enregisterment zeroing out some per-feature meaningfulness in the process of regimenting the meaning of entire repertoires.

Eckert’s (2012) statement on social meaning focuses on persona style, which she identifies with “enregistered voices” (Agha 2005) even though styles like “diva” and
“caring doctor” are not widely enregistered). This leads to a focus on style and repertoire as an attribute of the speaker, linked to the concept of “accent,” what Agha calls “a sound pattern linked to a framework of social identities” (2007a, 191). We can provisionally define the “voice” in a similar way, except that where the accent is only sensible in relation to the linguistic structure of a given variety, the voice’s relationship with language structure is less cozy. The voice, in Dolar’s (2006) words, “anticipates meaning” and precedes linguistic regimentation of the speech signal. Voice is also less “social” than accent, tied more directly to the individual speaker, and at least partially grounded in their physical reality. While, as Laver (1968) points out, it is commonplace to judge someone’s health by their “voice,” it would ordinarily be absurd to make similar judgments based on their “accent.”

2.5 Context-dependence

However much we want to talk about styles as they relate to persons or to personal social identities, it is also legitimate to talk about the stylistic profile of a particular text or performance, which means we will be using a lot of words like “context” and “co-text.” Agha (2007a) proposes two ways that indexicality can emerge in a text. The text-default associated with any particular meaningful stretch of text is the kind of indexicality associated with a sign in-and-of-itself. It is detachable from context—i.e., it works as an index with the same value in different contexts. As I interpret the term, it is, as of the time of interaction, equivalent to the form’s nth-order indexical meaning. If the form’s meaning arises historically through enregisterment, its text-default is the (always tentative) “endpoint” of the enregisterment process. In interaction, though, we typically encounter a particular richness of co-occurring indexical forms, giving rise to a second indexical modality: text-level indexicality. This is the indexicality achieved
by the multimodal deployment of multiple meaningful components in a single “chunk” of a semiotic encounter. The effects of text-level indexicality are “non-detachable” in the sense that they depend on the operation of all the individual semiotic components that make them up. The use of *wele* in the Lakhota example is a great example of text-level indexicality. The speaker’s deployment of the marker (whose text-default indexicality is simply “speaker=female”) would not function as indexing “concern” without the co-occurrence of indexes of his maleness—not just in the linguistic signal but also in the role configuration between him and his son, and other cues from the larger “context.” There are parallels between this picture of text-level indexicality and the “compositionality” of certain sociophonetic styles, both being, in a sense, packaged at the time of interaction and deriving their effective meanings from the interplay between their constituent features.

There is ample evidence that sociophonetic components of meaningful styles (both compositional and enregistered) achieve their effect only through the mediation of text-level indexicality—that is, through the active synergy of co-occurring signs. Some studies (Munson, McDonald, DeBoe & White 2006) have found that higher peak frequency of /s/ cues perceptions of male speakers as gay. Campbell-Kibler’s (2011) study of perceived sexual orientation confirms that this “fronted” /s/ carries stables stereotypical associations of more gay, less masculine, and less competent. But /s/-fronting interacts with another variable, (ING), to change percepts in complex ways. These results (see also Levon 2011 on intersecting variables in the Hebrew of gay Israelis and Pharao et al. 2014 on /s/ in Denmark) point out that the even strongly stereotyped variants like fronted /s/ can be modulated or canceled out in context.

The indexical field (Eckert 2008) helps us sort out some of this variability. We visualize (as in figure 2.3 and figure 2.4) an array of meaning potentials arranged according to relative similarity and ordered from more fleeting stances in the center to more
enduring personae at the periphery. In figure 2.3, the different labels are meanings that have been assigned to released /t/ (e.g. the [tʰ] in a “careful” pronunciation of [spitʰfaɪə] ‘spitfire’). These meanings, as Eckert points out, are not randomly occurring but are characterized by greater and lesser degrees of similarity to each other, as well as different degrees of perdurance (the timescale at which they are expected to hold). The indexical field maps out these similarities and other qualities. For instance, an “emphatic” reading of /t/ release is more likely to be re-interpreted as “annoyed” than, say, a “clear” reading is (and, in theory, much less likely than a “formal” reading is). The stance of “annoyed” is, meanwhile itself linked to a social category from a particular ethnographic context, “nerd girls” from Bucholtz’ (1999) ethnography of a California high school.

One issue that has incited some curiosity has to do with the ontological status of the indexical field. Where does it live? Does it inhere to the variant? Is it a property
of an individual’s mind? Maybe it is a property of a speech community or community of practice—or perhaps it reaches beyond the boundaries of individual communities. In my view it is most profitable to say “yes, but not necessarily” to all of these potentials. The indexical field is at root an analytical construct, a synoptic technique for visualizing the panoply of indexical meanings a form can take on in circulation through interactional and macrosocial timespace. It should be an aid to understanding how different meanings and functions for a variant arise, and how these meanings are linked to each other. At all times it must be put into dialogue with the traditional concerns of any variationist: what is the community (or “domain of evaluators,” in Agha’s analogous phraseology) under study? What is the structural “envelope” of the variable and its realizations? The space of language structures, ideologies, linguistic practice, and social differentiation that can structure indexical meaning is specific to particular ethnographic contexts. The indexical field, then, is equally constrained by ethnographic context. This is not the same as saying it is located at a particular “level,” such as the speech community, since it could equally well be adapted to describe the meanings of a form as it travels between speech communities. The indexical field concept, grounded in Silverstein’s firmly stated commitment to sociolinguistic analysis as a dialectic enterprise, is not, I believe, best taken as a real structural “entity” of any sort, but rather as a clearinghouse for attested variant meanings, arranged in an ethnographically validated space of relatedness.

Part of what helps us leap from lily pad to lily pad (and upward in indexical order) in this space of indexical meanings is “context,” or co-occurring indexical signs. It is often only given a favorable interactional context that a released /t/ presupposing “careful” can be used to perform “prissy.” This is certainly the case in the the indexical field diagram for tag questions in an English high school (figure 2.4 from Moore & Podesva 2009). There, the issue is about the use of tag questions, as in the example
“Ellie and Annabel don’t really get along, do they?” (2009: 465). Moore and Podesva argue that tag questions have a shared function: a “conducive” quality that invites or compels the interlocutor to join in some shared activity. What that activity is—for example, criticism of non-present others, involvement in co-constructing a narrative, and so on—depends both on which social group is using them and, necessarily, on the other signs and discourse processes the tag co-occurs with. We can see this in the case of the “Eden Village girls,” a small community of practice of girls who lionize a particular kind of hegemonic feminine appearance and demeanor, one which most Anglo-Americans will recognize—focused on well-manicured outward appearance, sexual restraint, and unimpeachable morals. When an Eden Village girl uses a tag question to increase “involvement,” she may also be interpreted as making a bid for increased “friendliness.” At least theoretically, these practices may be further enregistered, if perhaps for some small domain of evaluators, as indexical of group membership (i.e. “Eden Village girlness”) itself.

2.5.1 Relative Motivation

There is an implicit center-to-periphery motion in these indexical fields. In the examples from both /t/ release and tag questions, going from a more “central” meaning to one further out involves increases in indexical order—that is, the forging of creative meanings by filtering presupposed meanings through context and ideology. As the discussion of indexical inversion above establishes, however, it is not clear that hopping from center to periphery in the indexical field can always be characterized as an increase in indexical order. Rather, what seems to characterize the structure of those indexical fields that have been proposed so far (especially Eckert’s [2008] field for /t/ release and Moore and Podesva’s [2009] for tag questions) is degree of relative motivation.
To discuss this clearly, I should re-emphasize that conventionalization and “motivation” (or lack of arbitrariness) are not entirely dichotomous. The indexical senses of “clear” for /t/-release and “conducive” for tag questions are conventional as well as motivated. In the “conventional” column, ideologies of articulation as “full” pronunciation of segmental content and the function of questions as being ways to conduce interactional participation are both historically contingent.  

However, these “ideologies of the signal” construct an object which can then be evaluated in an iconic mode. The proposed “core” meanings of /t/-release (“clear,” “careful,” “emphatic”) and tag questions (“conducive”) are highly motivated relative

---

5 See Briggs 1986 for ethnographic examples of questions being a bad way to conduce others to action.
to these ideological objects. They pick out iconic links (associations based in resemblance) between the signal and the speaker’s stance (from “full” pronunciation to “clear,” “careful,” or “emphatic” stances, or from ideologies of the function of questions to a “conducive” stance that invites or coerces hearer participation). These links are cross-modal (from signal to stance), and grounded in resemblance—iconic.

Furthermore, as you hop away from the core in any contextually or ideologically driven direction, the arbitrariness of the relation between the sign and its proposed object increases. A meaning like “articulate” bears an ideologically motivated relation to clarity, but the ground of the relationship between articulateness and /t/-release itself is less motivated, as the ideologically structured “signal” provides no iconic ground for assessing a production as articulate or not. The increasing arbitrariness of meanings like “prissy” or “diva” or “gay” brings us to the edge of the field and into the domain of enregistered voices, where it is most apparent that a particular form-meaning relationship could have turned out entirely differently, given a different history.

Importantly, peripherality in this sort of indexical field is not equivalent to indexical order. The inverted indexical order of wa in late Modern Japan, for instance, goes from a presupposed (nth-order) indexicality of “woman” towards an $n + 1$st indexicality of “soft,” an interpretation grounded in the ideology of gender and language. But the relation between softness and womanhood has been essentialized, so that softness surfaces for interpreters as an iconized (in Irvine and Gal’s 2000 “ideological” sense) property of women talkers. In this way the “softness” of wa is naturalized and its role in the contingent history of Japanese femininity is erased.

The question of relative motivation, or iconicity, in stylistic meaning looms large at present. The concept of iconicity—sign-relations motivated by similarity between sign and object—cannot exhaustively determine the criteria in light of which one object
will be considered similar to another. It is worth being persnickety about maintaining that iconic sign-relations must also be filtered through culturally relative systems of interpretability (Irvine 2005). For example, some of the meanings proposed by third-wave variationists, such as “inarticulate” for “G-dropping” in (ING) variability, are straightforward iconizations. Based on an (objectively inaccurate) ethnotheory of the speech signal, which labels “G-dropping” as reductive ($<G>\rightarrow \emptyset$) rather than suppletive ($\eta \rightarrow n$), language users draw links between speech behavior, speaker demeanor and linguistic structure. As another example, Irvine & Gal (2000) describe how speakers of Nguni, a Bantu language in southern Africa, incorporated clicks into their language as a component of registers of forms for replacing tabooed lexical items. The clicks, borrowed from Khoi languages unrelated to Nguni and spoken by social “others,” were used because their “otherness” iconically resembled that of the Khoi speakers from which they were borrowed. Of course, the “otherness” of Khoi speakers for Bantu peoples is as good an example of an ideological phenomenon as any. In both of the above cases—G dropping and click iconization—iconically grounded sign relations (as performed by ordinary speakers and hearers) conflict with linguists’ and ethnographers’ expert reconstructions of “reality.” Given this conflict, it is relatively easy to keep the contingent and ideological nature of the iconization process in mind.

Sometimes, though, it is other linguists who provide the iconic grounds for interpreting stylistic variables, such as applying versions of the “Frequency Code” or “Effort Code” (Ohala 1983; Gussenhoven 2004) to intonational variation or creaky voice. These theories, which posit that gradient, natural properties of the acoustic signal are inherently linked to certain interpretations, do not “automatically” expose the operation of ideology as in the case of “G dropping” (where we get a convenient clash between folk theories and expert theories). Particularly when these codes are theorized as being hardwired into our cognitive circuits for interpreting “paralan-
guage” do the tools of ideological and historical analysis come to appear inapplicable—even when evidence from cross-language and cross-cultural variability in the interpretation of pitch and voice quality provides clear evidence that history must intervene at some point! So, however un-ideological the relationship between low speaking fundamental frequency and “large size” may seem, we must remember that this link, to the extent that it is even active or relevant, is also mediated by ideology at some point in the process.

A good example of the commingling of arbitrary social meanings and iconicity comes from Sicoli’s (2010) account of “voice registers” in Zapotec and throughout Mesoamerica. Voice quality in Zapotec is an indexical cue to the relative status of participants in a conversation. Falsetto voice is enregistered as the primary component of a register known locally as “respectful voice,” the text-default value of which is to indicate that the addressee is of higher status than the speaker. Having undergone enregisterment and serving a culturally relative and rather distinctive function, there is no question that this indexical meaning is conventionalized. However, as Sicoli points out, there are similarities between the use of falsetto in Zapotec and the characteristic high pitch of low-caste griots in Wolof society. In addition, in Sicoli’s data, the actual fundamental frequency of the “respectful” speech is also iconically related to the status of the addressee, with higher-status addressees eliciting progressively higher pitch implementations of falsetto voice. These two phenomena are evidence that iconism of some sort is at work in generating the social meanings of voice register tokens, according to some schema such as \([\text{status of addressee} \rightarrow \text{pitch}]\).

Sicoli explicitly calls for active critical engagement with Ohala’s and Gussenhoven’s ideas (which he identifies as motivated by “ad hoc reasoning” when confronted with exceptions to their principles). As he affirms, “we would do better to use ethnography and prosodic description together to make explicit the orders of semi-
otic process that mediate the patterned co-occurrences of voice qualities with speech contexts and social functions” (548).

Analyses of /r/-lessness and its social meaning also confront the iconism problem. /r/-lessness is variously enregistered as a feature of British English (for Americans), Received Pronunciation (or RP, especially for evaluators in the UK), the Boston accent, the New York accent, the Southern (US) accent, (US) black accents, and so on. In an analysis of style-shifting for (r) among Labov’s (2006 [1966]) New York city informants, Silverstein (2003) argues that the structure of the sociolinguistic interview highlights American standard language ideologies. At the “ritual center” of this event lies reading: the act of converting the written word to the spoken. When the grapheme ‘r’ does not faithfully appear in the spoken “output” of this process, Silverstein holds that its absence is ideologically interpreted as reduction. Given that /r/-less productions often do involve objectively reduced segmental productions and the neutralization of phoneme distinctions (in contrast to the phonetics of ING alternation), it seems unsettlingly easy to propose that the meaning of /r/-lessness is in some way iconically related to the idea of “reduction.” But this move would present some problems.

First among these is the position of /r/-lessness at the top, rather than the bottom, of the standard language heap across the pond, and indeed in most English-speaking locales outside North America. In England if not in the UK more broadly, /r/-ful pronunciations generally have low value in standard language markets, i.e. have low overt prestige, whereas RP has high overt prestige. But if /r/-lessness is as evident in “grapheme-to-phoneme conversion” in England as in the US, which it should be, why is it evaluated so differently?

First, I suspect that “reduction” does not even come into play in the field of meaning potentials for /r/-lessness across the pond. Why not? Enregisterment may
have a role to play. The shifting “cultural values” (Agha 2007) of Received Pronunciation as a prestige or standard phonology link the linguistic features of that register to a repertoire of performable personae. The intensity with which these associations circulate may be enough to countervail any thought of activating a latent iconic link between orthographic and phonological plenitude. Standard English in the US is a different beast altogether, not associable with a unitary accent (despite comparatively halfhearted appeals to a basically fictive “Midwestern” or “newscaster” accent), not phonologically well-defined in any case, and embedded less in the semiotics of class and region that have long operated in the UK and more in the semiotics of education, where reading as “ritual” may indeed have more significance. Whether or not this is the case, if Silverstein’s explanation of style-shifting in Labov’s data is to be believed at all, the iconic link between /r/-fulness and the orthographic standard must be understood as a conventionalized one which is not operative in all cultural contexts.

2.6 THE SOCIAL MEANING OF CREAK

Context-dependence and relative motivation are quite relevant in the social meaning of voice quality. Non-phonemic creaky voice, in particular, has long interested scholars of language’s social meaning. As early as 1968, Laver proposed that voice quality conveys indexical information about a speaker’s biological configuration, psychological state, and social status. These links are often non-arbitrary, but Laver & Trudgill (1979) make specific stipulations that their uptake of indexicality to discuss meaning in speech is to include arbitrary as well as non-arbitrary meanings.

Podesva (2007), investigating falsetoo phonation in the speech of young gay men in California, has analyzed creak as a lowering of the bottom of the pitch range. Creak also has proposed indexicalities of “male” and potentially “hyper-masculine”
in the UK (Henton & Bladon 1988), “tough” in mediatized portrayals of Chicano gang members (Mendoza-Denton 2011), “authoritative” in the speech of U.S. women (Yuasa 2010), and so on. All of these could potentially have iconized links to “lowness.”

Lowness does not necessarily always pertain to pitch, however. Brown and Levinson (1987, 267-268) link creak in Tzeltal to a stance of “commiseration” that arises from a “low speech energy” that “can implicate calmness and assurance.” Phonation type has also played a role in the areas of sentiment analysis and phonetic correlates of emotion. Gobl and Ní Chasaide link their category “lax creaky” to emotions with “low activation,” including boredom, which is also mentioned by Laver (1980, 126).

### 2.6.1 Creak, Iconism, Arbitrariness

Though some scholars, notably Gussenhoven (2004), have proposed theories of paralinguistic (mostly intonational) meaning as arising from biological design, such theories can at best sketch out a typology of the iconic “ground” available for signers to make sense of the meaning of something like creak. Many of the meanings proposed above share a quality of lowness, either of pitch or “energy.” This may straightforwardly index the speaker’s “low” affective state, as in the “calmness” of commiseration-seeking in Tzeltal, or the boredom of creak in English. More speculatively, the emotional restraint and commitment to silence indexed by creak in Chicana gangster Babygirl’s performance of “hardcore” demeanor in conversational narratives (Mendoza-Denton 2011) may also find an iconic ground in this kind of “low” activation.

Masculinity and maleness loom large in some of these meanings. Maleness may be indexically linked to “lowness” via widespread awareness of differences in male and female speakers’ mean speaking fundamental frequency, which is generally lower.
among men cross-linguistically, requiring an iconic leap from “low pitch” to “low speaking energy.”

Meanwhile, values of creakiness that have ideological links to masculinity (e.g. professionalism or authoritativeness), as in Yuasa’s work, may operate at one degree of historical remove from this creak→low pitch→male schema—hence their appeal to “upwardly mobile” women, and possibly indicating some degree of divergent conventionalization. Mendoza-Denton (2011) and Podesva (2010) argue, however, that by no means does gender necessarily play a role in creak’s semiotics.

Of course, there is still room for work on the more plainly arbitrary indexical values that creaky voice can take on. Mendoza-Denton (2011) describes specific mutations that creaky voice’s semiotic value undergoes across steps in an “indexical chain,” and Chun and Podesva (2010) highlight the mutability of other phonation types’ meaning in context and within the microdynamics of interactional time. This adaptability is part of what makes creaky voice so interesting—it serves many purposes.

In multiple social contexts, the fact that creak lies “below” the typical range of fundamental frequencies for modal voice (see Podesva 2007 and Mendoza-Denton 2011 on creak in English, as well as Sicoli 2010 on breath and falsetto in Lachixio Zapotec) is important in explaining its capacity for generating certain social meanings. Mendoza-Denton (2011, 265) and Podesva (2007) both argue (as does Sicoli, for breathy voice) that creak is sometimes meaningful as an expansion beyond modal pitch range. In Podesva’s data from self-recordings of gay men in California, this prosodic expansion could generate meanings of “animated.” For Mendoza-Denton’s narrative data from ethnography among Chicana gangsters in a California high school, creak responds “to a larger conception of how to manage one’s affect that includes not showing emotion” (2011, 265). These semiotic effects are almost perfectly opposite, but in both cases,
it is creak’s lowness (beyond the bottom of the modal pitch range) which enables its contribution to the semiotic whole.

2.6.2 **Creak’s context-dependence and multiple meanings**

The contrasts and parallels between creak in the contexts which Podesva and Mendoza-Denton describe highlight that the factors that structure which of creak’s meaning potentials get activated is not fully understood, nor is what those meaning potentials even are. We can find conventionalized iconic grounds to begin theorizing creak’s links to meanings such as “boredom,” “hardcore,” and “animation.” But what about “maleness,” “masculinity,” or even “sexiness” ([Kajino & Moon 2011](#))? Eckert’s (2008) proposal for the “indexical field” implies that some kind of ideological or interactionally achieved similarity in a meaning space is good enough to construct a path from any one meaning to the meanings closely related to it, but it seems to me that this imposes at once too much structure (why only two dimensions?) and too little (how much proximity is enough, and how do we determine it?). As Campbell-Kibler (2007) highlights in her argument for the role of “accent” in mediating-listeners’ evaluations of a variable, the packaging of social attributes and linguistic features may be structured in a somewhat lumpier or more hierarchical fashion than the indexical field currently highlights. In addition, the kinds of factors that can intervene in structuring “indexical variation” are quite various: social divisions (as in the case of community of practice or class divisions in a speech community), co-occurring linguistic signs, other co-occurring signs in other modalities, speaker experience ([Johnstone & Kiesling 2008](#)), language ideologies and, as I argue in this dissertation, language structure (though see also [Silverstein 1976, 1979](#) for related arguments about language structure). We should remain alert to the fact that these factors may affect variability indexical meanings in different ways.
In any case, we can be fairly certain that “context,” taken to mean both co-textually present indexical signs and more perduring metapragmatic ideologies, has a key role to play in constraining meaning variability. One factor that Campbell-Kibler (2007) and Podesva, Jamsu, Callier and Heitman (n.d.) have highlighted is the hearer’s preconceptions about the talker. This can include conclusions about their social identity, as in Campbell-Kibler’s work—where talkers rated as “Southern” were heard as more accented when they used alveolar guises of (ING), but talkers rated as “gay” were more accented in their velar guises of the same variable.

Podesva (2010) proposes an indexical field for creaky voice based on previous literature and his work on creak in Washington, DC (see figure 2.5). One goal of this research will be to sketch out a similar field for creaky voice in Mandarin, exploring the meaning potentials of creaky voice. In so doing, though, I also hope to pay heed to creak’s participation in stylistic repertoires, both enregistered and emergent.
2.7 Creak in stylistic contexts and co-texts

In a study of two speakers, Kajino and Moon (2011) found that creaky voice could be associated with sexiness, but only in one speaker’s style, helped along by the other elements of her performance. Podesva (2007) finds increased rates of creak alongside increased rates of falsetto voice. Ogden (2004) proposes a hierarchical ordering characterizing sentence-final non-modal voice; while creak is the most common sentence-final non-modal voice, it can also co-occur with breathiness, whisper, voicelessness and exhalation.

Farris (1995) reports that the petulant, whiny sajiao style of Mandarin Chinese includes pitch and voice quality variations that are accompanied by increased affective particle use and prosodic modulation. While not all styles are identifiable by names like sajiao is, it would be interesting to investigate what kinds of indexical sign configurations appear commonly with creaky voice.

In my own work, I explored the use of various voice qualities on a Chinese TV drama (Callier 2012). Creak was used by several characters but only one character used it often. As a middle-aged widow in love with the protagonist’s main sweetheart, as well as a shrewd business operator, this character often finds herself pitted against both established authority (the primary villains) and the protagonists, who are best described as young, plucky upstarts. Her character (who I have previously classified as a “villain” but is ultimately more complicated) negotiates a tricky dialectic between public obligations and personal self-preservation, and her creaky voice may be related to her depiction as an embattled, private, and highly practical person, as well as a threat to established orders of gender, sexual propriety and socioeconomic status. There are perhaps ties here to the affective restraint of Mendoza-Denton’s informants, but the specifics of this situation remain to be explored in more depth.
2.8 Prospectus

In this dissertation, in particular [chapter 5], I will be investigating the indexical meaning potentials of creak. Specifically, I will be assessing creak’s meanings in different linguistic contexts, to see if the linguistic environment a variant occurs in does anything to inflect its social meaning. The literature does not, a priori, point to any specific mutation being more likely than another. But in accordance with the general principle that the deeper one looks, the more variability there is to find, I do indeed find that linguistic context changes the social meanings listeners assign to creaky phonation.
In this chapter, I will use a corpus of Mandarin sociolinguistic interviews collected in summer 2012 to investigate some research questions about the distribution of creaky voice in Mandarin Chinese: are there social differences in frequency of creaky voice? Does creak occur more in particular tonal or prosodic environments?

If Beijing is akin to the British context that Henton & Bladon (1988) studied, we would expect creak from males, possibly with a particular regional distribution; if it is like Yuasa’s (2010) contemporary California, we would expect creak to be used more by women. As for linguistic contexts, based on prior work (Belotel-Grenié & Grenié 1997, Keating & Esposito 2006) we expect tone 3 (low dipping) and tone 4 (high falling) to exhibit creak, as well as IP-final contexts. As we will show, there is no appreciable social distribution of creak, and its linguistic distribution appears to follow the patterns described in the previous literature.

3.1 Participants, data and methods

This section will step you through how I put together and analyzed the corpus of speech production data that will help answer the above questions, addressing recording environment and sound, transcription and annotation, forced alignment of the speech data, acoustic measurements, and building quantitative models.
3.1.1 Recording and participants

I performed a number of sociolinguistic interviews (semi-structured interviews covering topics designed to promote conversation and interviewee interest, see Labov 1972b, 2006 [1966]) in 2012 in Beijing, China, with the intent of studying creaky voice. I was affiliated with Beijing Normal University (BNU) at the time, and my interviews generally covered family background, childhood games, educational history, and work history (if applicable), as well as, depending on the interviewee: hobbies, plans for the future, and media consumption habits. I also asked interviewees to talk about media figures or people in their lives whose voices they found unique or interesting. Interviewees were sometimes interviewed alone and sometimes in groups of two or three; interview encounters lasted between a half-hour and one hour.

The final collection of interview recordings included 20 speakers: 7 male and 13 female. Most participants were contacted through a vocational school in suburban Beijing, while 2 others were identified through my contact network. All were born, raised, and lived most of their lives in the Beijing area, but some were from less urbanized outlying districts while others grew up in the more urban inner districts (see figure 3.1). There seemed to be salient differences in the kinds of experiences, interests, and practices that “urban” and “rural” speakers had, so I included urban/rural status as a factor in analyses in this chapter and chapter 4. I describe how I made this distinction in subsection 3.1.2.

Some speakers were not included from the total collection of recordings: two female speakers (the ones identified through my informal contact network) were excluded due to dramatic difference in age from the rest of the sample, while two further speakers (one male, one female) did not contribute enough speech to be analyzed. The 16 participants that are analyzed in this chapter are listed in table 3.1 along
Figure 3.1: From Beijing Tourism Group n.d.
Table 3.1: Participants with demographic information and number of vocalic segment tokens used for analysis.

with their sex, rural/urban affiliation and the number of vocalic segments extracted from recordings of their speech.

Uncompressed PCM recordings were made with an Olympus LS-10 in as quiet a location as could be identified in each encounter. In the vocational school, this was an empty office. I transcribed portions from the beginning of each recording, segmenting the audio first by IP. The quantity of speech transcribed varied for each speaker, but the “language” modules (questions about language characteristic of Beijing and unusual voices) were never included (see Labov [2001]).

The school where I did the interviews analyzed in this dissertation was days away from letting out for the summer when I arrived, and is located in a dusty suburb.
of Beijing connected to the rest of the region by buses as well as a recently opened subway line. It was a vocational school, and the students I spoke to were training for positions in various sectors of the state-controlled economy.

The conversations I had with informants were often very stilted (at least as far as I perceived them), for various reasons. They took place in administrative offices in the students’ academic unit; being so close to their teachers’ offices may have put them on their guard. Furthermore, I would describe my Mandarin proficiency as “conversationally fluent,” but certainly far from native-like. Often interviewees found themselves having to repeat unfamiliar words for me, explain talk that I hadn’t understood, or wait for me to formulate a question or prompt. The discourse characteristics of these “barriers” to “smooth” communication—which were a bigger obstacle in some interviews than others—will have to await investigation in another venue, but they bear mentioning.

Qualitatively, I noted a wide range of vocal qualities across participants, as well as accents and other distinctive linguistic traits, when I was talking with my participants. Yongxiang and Weihong, two male friends from urban Beijing who came into the interview together, had a classic Beijing accent that I can only describe as “crisp.” Yongxiang is one of the creakiest speakers in the corpus, and Weihong is also fairly creaky. On the other hand, Dong’e, who was interviewed along with three other female friends (not analyzed here), was teased about her “Changping” accent, Changping being the mostly rural outlying suburban district from which she hailed (and where the school was located). Dong’e struck me as a very quiet speaker, but other than a relative lack of “crispness” compared to the likes of Yongxiang and Weihong, without proper linguistic analysis I could not have immediately put my finger on what made her accent distinctive. She and her friends did cite the use of a locative marker _generic in Changping, which according to them is not used in the urban Beijing variety (but
is homophonous with 跟 gèn ‘with’). Some interviewees showed signs of having been
told that I was interested in language in Beijing; Wenhan, for example launched into
descriptions of the vanishing voices of roving vendors hawking their wares in Beijing’s
old alleyways shortly after I asked his name. Some interviewees—Dong’e was a good
example—were also very comfortable asking me questions, for instance about where
I was from, where I was staying in Beijing, and so on.

My own subjective impressions are of course informed only by relatively brief
encounters with the speakers and shallow ethnographic involvement. Though by then
I had several years of a Mandarin language instruction under my belt and about a
year’s experience living in Beijing, I had much less impression of what constituted,
for instance, a “Changping” accent, or any of the other nearby regional accents, than
I did of the “proper” Beijing accent.

It may be worthwhile to note that certain forms of linguistic diversity were
extremely well-known to me but evaded comment by any of my informants. The
voices of migrants from more far-flung Mandarin-speaking regions, or from other
dialect/language areas entirely, are quite common in Beijing, and linguistic insecurity
among migrants to the city is high, alongside persistent economic uncertainty and
social and political marginalization (Dong 2011). The “cosmopolitan Mandarin” of
Zhang’s (2006) investigation of the speech of yuppies in Beijing’s private “market”
economy was also “in the air” in Beijing, mostly as “Taiwanese-sounding” accents
in mass media, but practices associated with it (avoidance of neutral tone, English
code-mixing, etc.) were never mentioned when I asked informants to talk about lan-
guage in Beijing. As participants were prospective participants in the state-controlled
economy, this might be as expected. In addition, the “Beijing dialect” is so strongly
enregistered that any invocation of Beijing as a linguistically meaningful place seemed
to invite talk about Beijing dialect.
My orientation in this dissertation is mostly toward phonetic details of my
speakers’ talk, mostly in the aggregate or across certain structurally determined
categories. I am unable, in this relatively short space, to elucidate their discourse
goals, present the identities fashioned for them in interaction or which they wanted
“the world” to see, or meaningfully address the performative impact of their talk in
our interactions. Furthermore, in looking at social variability in voice quality among
these speakers, I am not making strong assertions about the local validity of sex or
“region” as meaningful categories for my informants. Nevertheless, as a means of
generating hypotheses for ethnographic engagement or for further variationist work
(or both), it is definitely worthwhile to ask questions about creak’s social distribution.
Establishing the categories used in coding social and linguistic factors was not always
simple, however.

3.1.2 Coding considerations

Sex was coded as a binary variable with values ‘Male’ and ‘Female,’ based on my own
impressions and interaction with the participants. Region was also binary, taking
values of ‘Urban’ or ‘Rural.’ Though the dynamics of Beijing geography are far too
complicated to reduce to a single dichotomous variable, I chose this divide to reflect
apparent divisions in my speaker sample in terms of many factors, including where
leisure time was spent (in peripheral commercial areas or downtown Beijing malls)
and what sort of building and community speakers grew up in (courtyard houses or
apartment buildings more common in urban Beijing, versus villages, e.g. “村儿” cunr
‘village’, etc.).

The urbanization of the Beijing area has generally proceeded outward from the
traditional core (Xicheng, Dongcheng, Xuanwu and Chongwen), through the closer
’suburbs’ of Fengtai, Haidian and Chaoyang (largely urbanized as of the late 20th
century), and onward. Today, following the swift development of the Beijing area, significant portions of Shunyi, Changping, and Tongzhou are also highly dense, urban in character, and are connected to the rest of the Beijing conurbation by trains, public transit, and high-capacity roadways. In particular, the “ring roads” of Beijing, multi-lane freeways that encircle the city with the old city core roughly at the center point, have become emblematic of the outward expansion of urbanization. They are often referred to in conversation and public advertising for new developments, and are almost unavoidable in cross-city automotive transit. While it is problematic to use these, or any other fixed boundary, as reference points in determining rural/urban divides—given the rapidity of urban expansion—I decided to use one of the ring roads as a provisional dividing line. Speakers who originated within Beijing’s fourth ring road were coded as Urban; all others were coded as Rural. Although today it is the fifth ring road, and in some quarters, the sixth, that serves as the best “line” between the city and its surroundings, at the time when these speakers (most of them around 20 years old in 2012) were growing up, urbanization had not proceeded quite so far (Ma et al. 2005).

IP Position is a continuous variable calculated for the $n$th syllable in an IP of length $N$ syllables as $(n - 1)/(N - 1)$. Tone was a nominal variable that took on five values: 1 . . . 4 for the four lexical tones of Mandarin and ‘5’ to represent the ‘neutral tone’ of lexically destressed or inherently toneless syllables. Tone 1 is a high level tone (55 in Y.R. Chao’s 1965 tone numerals), tone 2 is a high rising tone (35), tone 3 a low dipping contour (21[4]), and tone 4 a high-falling tone (51). In most figures and discussion below, I will refer to the neutral tone as tone 5, but this does not necessarily mean I regard it as a full-fledged tone category on par with the other 4.

Since the frequencies of the first few formants can strongly impact spectral tilt measures of phonation (Esling & Edmondson 2011), it is important to control for
vowel identity in analyzing spectral tilt measurements. But what vowel categories to use? Most descriptions of Mandarin vowels start from a description of their distribution that includes at least 11 allophonic categories. Phonemic analyses of these usually posit a far lower number of categories, from three (Hockett 1947), to five (Huang 1992; Duanmu 2007) or six (Cheng 1973) vowels. The most common descriptions of Mandarin recognize the allophonic variants shown in table 3.2, which assumes Duanmu’s five-vowel system.

This analysis does not cover the full range of variation in Mandarin vowels. For instance, Duanmu (2007) reports that /ɨ/ approximates [u] in certain tonal environments, and some vowels have glide realizations (e.g. /iə/ > [je]), or lax variants that appear in diphthongs (/tʂau/ > [tʂu]). Recognizing that the impact of vowel class on measures of spectral tilt depends mostly on vowel height and not on specific

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Allophone</th>
<th>Environments</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>i</td>
<td></td>
<td>xi [ci], di [ti]</td>
</tr>
<tr>
<td>i (1)</td>
<td>after dentals</td>
<td></td>
<td>si [si], zi [tsǐ]</td>
</tr>
<tr>
<td>uy (1)</td>
<td>after retroflex</td>
<td></td>
<td>shi [ʂu], zi [tʂu]</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
<td></td>
<td>ju [tɕy]</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td></td>
<td>du [tɕ], ku [kʰu]</td>
</tr>
<tr>
<td>ø</td>
<td>ø</td>
<td>in open syllables, except after labials</td>
<td>nen [nən]</td>
</tr>
<tr>
<td>ø</td>
<td>ø</td>
<td>after labials</td>
<td>de [tɕ], ke [kʰu]</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td></td>
<td>bo [po]</td>
</tr>
<tr>
<td>e*</td>
<td>after /i/</td>
<td>ye /iə/ &gt; [je], die [tje],</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td></td>
<td>da [tɑ]</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>before velar nasal finals, or before or after back vowels</td>
<td>gang [kɑŋ], zhao [tʂau]</td>
</tr>
<tr>
<td>e*</td>
<td>/i_ɨ/</td>
<td>tian [tʰiɛn]</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Five-phoneme analysis of Beijing Mandarin, with allophones (allophones that appear only as nuclei of diphthongs are marked with an asterisk)
vowel class (Esling & Edmondson 2011), I will use the above list of the segmentally conditioned allophones of Mandarin Chinese vowels for the random effect of vowel segment class in the following analysis.

3.1.3 Transcripts and forced alignment

Transcriptions were made in simplified Chinese orthography using ELAN (Max Planck Institute for Psycholinguistics, The Language Archive 2012) and then converted to the pinyin romanization system using the Adso Chinese Translation and Annotation Engine (Lancashire 2011). Based on cursory inspection, these automatically produced romanizations were 95-99% accurate, with most errors the selection of a phonologically similar homograph. Because of these high accuracy rates, I did not hand-correct any of the romanized transcripts except where issues could be resolved by improving the original transcript. Because the romanization provides a basically phonemic guide to pronunciation, I was able to use the romanized transcripts to perform “forced alignment” on the audio, a technique that uses speech recognition to align a sequence of phonetic labels with timestamps in the audio signal. This yields a large quantity of time-aligned phonetic data.

The forced alignment system used was the University of Pennsylvania Phonetics Lab Forced Aligner (P2FA, Yuan & Liberman 2008). The system takes a text transcript, converts it to a broad phonetic representation, and uses acoustic models of the phones (in this case, trained on a corpus of US Supreme Court oral arguments in English) to find the most likely alignment between the phone sequence and the audio signal. I used the P2FA system mostly unmodified, adapting only the dictionary that maps words (in this case, romanized pinyin representations of individual syllables, without tone marking) to phonetic strings. Although the segmental phonetic inventories of Mandarin and English are very different, I attempted to find a
“close enough” mapping such that the available acoustic models would substantially match the intended phones in the audio signal. The mapping I devised is described in appendix A. In a spot check I did of 100 segments per speaker, the majority of vowel segment alignments were correct and within 50 ms of an optimal position (“Match” in table 3.3). Smaller percentages were misaligned slightly but still contained the target segment (“Partial Match”), and some were misaligned entirely (“Mismatch”). Though I did not hand-correct the alignments, the issue of their reliability is important to address.

Table 3.4 depicts which categories of segments were most often misaligned. The category “silence” includes portions where the transcript indicated silence. Silence could be introduced for a pause or for gaps between phrase segment boundaries and the beginning or end of speech. Silence could also be the reflex of a “short pause” (sp) token inserted by the automatic aligner where there are brief silences. Silences has a high probability of misalignment, as did the scant tokens of noise (including the aligner-recognized tokens ns for noise and lg for laughter). Regression analysis (table 3.5) confirms that silences are less likely than obstruents, vowels, and sonorant consonants to be correctly aligned. Since only vocalic tokens are analyzed in this chapter, this provides significant support for the validity of the methods adopted here.

It is important to emphasize that the goal of using forced alignment to analyze large amounts of phonetic data is not, at least at first, to obtain precise annotations of phone timing for every individual phone, but rather to economically create large databases of phonetic data that collectively serve as a good resource for estimating underlying central tendencies of whatever features are of interest. With this limitation in mind, it is important to ensure that the resulting corpus is big enough, and that
the influence of outlier measurements from mistaken alignments is minimized. These issues will be addressed below.

### 3.1.4 Acoustic Measurements

The measurements I make in this chapter are in the family of measures of “spectral tilt” or spectral slope, and are also known as “harmonic differentials,” because each one is a difference between the amplitudes of two harmonic peaks in an FFT spectrum. H1, H2, A1, A2, and A3 are the amplitudes (in decibels) of the first harmonic, second harmonic, and the harmonics whose frequencies are closest to F1, F2, and F3, (the first three resonances of the vocal tract) respectively. Figure 3.2 illustrates the locations of these measurements for an example spectrum. I took measurements of F1, F2, F3, F0,
Table 3.4: Percent correct alignments by type of segment

<table>
<thead>
<tr>
<th>Segment</th>
<th>Match</th>
<th>Partial Match</th>
<th>Mismatch</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>90.97</td>
<td>0.00</td>
<td>9.03</td>
<td>144</td>
</tr>
<tr>
<td>Noise</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
<td>2</td>
</tr>
<tr>
<td>Obstruent</td>
<td>73.65</td>
<td>3.38</td>
<td>22.97</td>
<td>296</td>
</tr>
<tr>
<td>Silence</td>
<td>42.13</td>
<td>0.00</td>
<td>57.87</td>
<td>235</td>
</tr>
<tr>
<td>Sonorant</td>
<td>74.38</td>
<td>4.96</td>
<td>20.66</td>
<td>242</td>
</tr>
<tr>
<td>Vowel</td>
<td>76.40</td>
<td>7.87</td>
<td>15.73</td>
<td>534</td>
</tr>
</tbody>
</table>

Table 3.5: Estimates and errors from regression model of linguistic and social factors on correct alignment. The intercept represents the expectation of the log-odds of correct alignment for obstruents coming from male rural speakers.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.0589</td>
<td>0.1571</td>
<td>6.74</td>
<td>0.0000</td>
</tr>
<tr>
<td>Glide</td>
<td>1.2757</td>
<td>0.3196</td>
<td>3.99</td>
<td>0.0001</td>
</tr>
<tr>
<td>Silence</td>
<td>-1.3718</td>
<td>0.1877</td>
<td>-7.31</td>
<td>0.0000</td>
</tr>
<tr>
<td>Sonorant</td>
<td>0.0179</td>
<td>0.1984</td>
<td>0.09</td>
<td>0.9281</td>
</tr>
<tr>
<td>Vowel</td>
<td>0.1355</td>
<td>0.1671</td>
<td>0.81</td>
<td>0.4177</td>
</tr>
<tr>
<td>Male</td>
<td>0.1742</td>
<td>0.1265</td>
<td>1.38</td>
<td>0.1684</td>
</tr>
<tr>
<td>Urban</td>
<td>-0.1994</td>
<td>0.1275</td>
<td>-1.56</td>
<td>0.1177</td>
</tr>
</tbody>
</table>

H1-H2, H1-A1, H1-A2, and H1-A3 for vocalic segments of each speaker. Measurements were taken at 5 equidistant points (‘chunks,’ below) in each segment in each syllable, though only measurements from vowel segments were used. Figure 3.3 shows the hierarchical structure of the corpus and visualizes the location of the measurements.

Recall that H1-H2 and (especially for high vowels) H1-A1 are measures of “short-range” spectral tilt, since the harmonics whose amplitudes they use are quite close together. H1-H2 is correlated with the “open quotient”—the amount of time that the glottis is open during an open/close cycle. H1-A1 has been associated with the
Figure 3.2: FFT spectrum of Peter Ladefoged pronouncing schwa [http://www.phonetics.ucla.edu/course/chapter1/vowels.html]. The harmonics whose amplitudes provide values for H1, H2 and other relevant measures are marked.

bandwidth of the first formant. “Mid-range” measures of spectral tilt—H1-A2 and H1-A3—have also been used to differentiate voice qualities, but the articulatory measure they have been most closely associated with is the velocity of closure—i.e., how quickly the glottis goes from open to closed when the open phase ends. The shorter opening time and higher velocity of closure in creaky voice are theorized to add energy to higher harmonics, easing the drop-off in energy in successive harmonics and resulting in more negative values of the spectral tilt measurements.

The algorithm the Praat script (acquired from Vicenik n.d. and superficially modified) uses for making these measurements is the following:
Figure 3.3: Hierarchical structure of corpus. $\sigma$ denotes the syllables composing the IP; c1-5 are the 5 equidistant points at which measurements were taken in each vocalic segment.
1. Resample sound (the target vocalic segment with 250 ms of padding on either side, no windowing) to 16000 Hz.

2. Extract formants using a window length of 25 ms, 5 formants, and maximum formant value of either 5000 Hz if speaker is male, or 5500 Hz if female.

3. Use formant tracker to track the first 3 formants in the sound, using different formant reference values according to whether speaker is male or female: $F_{1_{\text{ref}}} = 500$, $F_{2_{\text{ref}}} = 1485$, $F_{3_{\text{ref}}} = 2475$ for men, $F_{1_{\text{ref}}} = 550$, $F_{2_{\text{ref}}} = 1650$, $F_{3_{\text{ref}}} = 2750$ for women. Frequency and and bandwidth costs are 1.0/kHz, octave cost 1.0/octave.

4. Extract pitch contour with floor of 60.0 Hz and ceiling of 350.0 Hz, time step of 12.5 ms (the default).

5. Divide the target sound into 5 chunks of equal duration. All subsequent steps are to be repeated for each chunk.

6. Extract and record the tracked $F_1$, $F_2$, and $F_3$ values at the midpoint of the chunk.

7. Extract and record $F_0$ at midpoint of the chunk. If undefined, record no further measurements for this chunk and move to the next one.

8. Calculate an FFT spectrum on the chunk, extracted from the original sound using a Hanning window.

9. To find $H_1$, search within a frequency window of $F_0 \pm F_0/10$; $H_1$ is the maximum amplitude within this window.

10. $H_2$ is the maximum amplitude in the frequency window $(2 \times F_0) \pm F_0/5$. 

76
11. A1, A2, A3 are the maximum amplitudes in the frequency windows $F_x \pm F_x/10$, where $x$ is 1, 2 or 3 respectively.

12. Record the values for H1-H2, H1-A1, H1-A2, and H1-A3.

This algorithm identifies H1 and H2 in windows around F0 and $F_0 \times 2$, so it is sensitive to poorly tracked pitch, as well as (obviously) to poorly tracked formant values. Also, it produces no result for sound chunks where Praat does not find F0. Considering that some researchers (Yu & Lam 2011) have operationalized creaky voice as where the Praat F0 tracker fails, this could undercount creaky syllables, particularly those that involve a full break in voicing (which is common in this data).

Another issue is ensuring that we have enough data from each participant. As an exploratory measure to see whether we do have enough data, I sampled (with replacement) from the values of the spectral tilt measures (for vowels only) to create a large number of samples, of various sizes. Figure 3.4 shows the sample means for each of these subsamples, for each spectral tilt measure, for one speaker.

Figure 3.4 shows that, given at least a certain sample size, sample means tend to fall in a similarly sized band of values, with negligible changes from further increases in sample size. The critical $n$-value beyond which variance in sample means seems to bottom out is somewhere between 100 and 200 samples (each sample is, recall, one measurement from one ‘chunk’ of a vowel segment). This pattern was repeated for the other speakers in the sample. Figure 3.5 shows the standard deviation of sample means at a given $n$ for each speaker. For all speakers, the “elbow” of the curve where standard deviation begins decreasing at a vastly slower rate is around $n = 200$. This indicates that per-speaker sample means in samples of at least 200 begin to approximate the true mean of the speaker for each measurement.
Figure 3.4: Sample means for spectral tilt measurements taken at 20% intervals of vowel segments, speaker Haiyu, with increasing sample sizes. The light shaded band shows the region in which 95% of the sample means fell at a given sample size.
Figure 3.5: Sample standard deviations for sample means of spectral tilt measures on each speaker as sample size increases (all speakers). Note that different speakers have different amounts of data available.
These data are highly unbalanced. Not only do individual speakers contribute varying amounts of speech to the overall corpus, but because of the inherently uneven distribution of lexical types and phonological categories in that speech, there are several factors which need to be corrected for when assessing the results quantitatively. Whereas ordinary least squares regression, which helps account for the random nature of the observed variables and can provide some information on the probability that apparent effects are or are not due to chance, would be adequate in many analytic contexts, we need a technique that can account for the lack of balance in these data.

Mixed-effects regression modeling has come into widespread use in psycholinguistics (Jaeger 2008) and sociolinguistics (Johnson 2009) and it addresses, among other issues, the problem of unbalanced data. A traditional regression equation might relate the response datum $y_i$ to the predictors $x_{0i} \ldots x_{Pi}$ with an equation such as the following:

$$ y_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pi} + \epsilon_i $$

where $\beta_0$ is the zero-intercept of the response (akin to a mean value), $\beta$ is a vector of coefficients of length $P$ and $\epsilon_i$ is the residual “error term” (i.e., the remaining difference between the product of the coefficient and the data value and the observed response. Fitting a model such as the above typically involves finding values for $\beta_0$ and $\beta_1 \ldots P$ that satisfy some desiderata, for instance minimizing the square of the sum of the residual errors (the “ordinary least squares” method).

The coefficients in $\beta$ specify the effects of the independent variables, which are known as “fixed effects” because the levels of a discrete independent variable are

---

1This is not the only reason to adopt the modeling techniques chosen here, but it is a good one. See Johnson (2009) for a fuller exploration of these reasons.
understood as an exhaustive sample of all the possible values that factor could have
taken on. A mixed-effects model allows other terms to be added to the equation rep-
resenting “random effects,” which are factors whose levels are understood to have
been randomly sampled from a larger population (Baayen 2008; Johnson 2008). In
linguistic work random effects commonly include categories such as speaker (since we
work with small speaker samples from larger populations of interest), word/lexeme (since the data include small subsets of a language’s “full” lexicon), and other ident-
tifiable sources of variability not explicitly built into the study design as variables of
interest.

The terms added to the regression equation for the random effects can represent
adjusted intercepts for different levels of a factor (random intercepts), as well as more
complex adjustments for the slope of a regression line between the response and some
other variable across levels of the random factor (random slopes). Below, I use only
random intercepts.

The possible sources of random variability affecting measurements which this
design encounters are numerous. Speakers may have idiosyncratic voice qualities
(which, from a different analytic standpoint, might be interesting in its own right,
cf. Johnstone 1996), or individual lexical or segmental environments might have
their own impacts—either from the segment that the initial measurement was drawn
from or from the preceding or following environment. In addition, repeated mea-
surements (the “chunks”) were taken from each segment and each syllable. Both the
non-independence of measurements from within a segment and possible variability
from the specific position the measurement was drawn from (which “chunk”) should
also be taken into account. For these reasons, I include random intercept effects for
Speaker, Syllable (levels of which correspond to segmental phonetic ‘spelling’ of each
syllable, without tone), Segment (the vowel category from which the measurement is
Besides random effects, variable selection (for fixed effects) is a further issue to be addressed. In the interest of parsimony, it is generally desirable to choose a model with as few parameters as possible, and choosing fewer predictors avoids issues with possible collinearity between predictors. In this analysis, I choose models by selectively stepping-down from a “full” model, which has fixed main effects for Gender, Region of origin, Tone, and IP Position, as well as an interaction effect between Tone and IP Position. Models also specify random effects for Speaker, Syllable, Segment and Chunk. Starting with this model, I subtract each individual fixed effect in isolation from the others, creating a set of “one-down” reduced models with one fixed effect missing. If the Akaike Information Criterion (AIC, see Burnham & Anderson 2002) for any of these models is lower than for the full model, then the reduced model with the lowest AIC is chosen and the process is iterated by removing another fixed effect. If the “full” model has lower AIC than all the reduced models, the process stops. If this never happens, the process stops when the only fixed parameter is the model intercept $\beta_0$.

The \textit{lmer} function in R’s \textit{nlme4} package estimates the parameters of mixed effects models by a process called restricted maximum likelihood (REML). Some authors (Baayen 2008) have assessed the statistical significance of fixed effect estimates in mixed-effects models with $t$-tests. The degrees of freedom for such tests are calculated by subtracting the number of fixed effect parameters from the sample size. This method of estimating the degrees of freedom is remarkably permissive, as mixed-effects models have many (in our case, hundreds) more parameters than just those for the fixed effects, but there is no widely accepted corrective for this—at least, none that results in a test statistic. Below I will report the
p-values resulting from $t$-tests performed in this way, but the significance judgments they entail should be taken with a grain of salt.

Another way to assess whether the fixed effect estimates are believable or not is to use Markov chain Monte Carlo (MCMC) estimation (Kruschke 2010). MCMC estimates the model parameters many times over on small, randomly drawn subsamples of the data. By examining the range in which these estimates fall, we can answer what range of values the parameter might believably take. In this chapter, I report the 95% highest posterior density interval (HDI) for all fixed effects parameters, which describes the region in which 95% of the parameter estimates fall the most densely. If the 95% HDI for a parameter includes zero, we can say that zero is a believable value for that parameter (most likely meaning that the process the parameter represents has no effect—similar to failing to disconfirm the null hypothesis). Herein, I will treat as credible an effect when its 95% HDI excludes zero (and the term credible can be interpreted as having this meaning).

3.2 Results

The distribution of tonal environments is shown in table 3.6 and figure 3.6 shows the distribution of segments across the length of utterances, split up by tone.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1 (hi level)</td>
<td>1381</td>
</tr>
<tr>
<td>Tone 2 (hi rise)</td>
<td>1385</td>
</tr>
<tr>
<td>Tone 3 (lo dip)</td>
<td>1607</td>
</tr>
<tr>
<td>Tone 4 (hi fall)</td>
<td>2499</td>
</tr>
<tr>
<td>Tone 5 (neutral)</td>
<td>968</td>
</tr>
</tbody>
</table>

Table 3.6: Number of vocalic segments by tonal category
Figure 3.6: Violin plot of tone by utterance position (width of the ‘violin’ indicates the density of data points at that value). In accordance with the nature of destressing in Mandarin (which is usually domain-final), neutral tone segments are unlikely to appear in initial position.

The four models that were created according to the variable selection process outlined in 3.1.5 are described in tables 3.7, 3.8, 3.9, and 3.10, one for each measurement under consideration (H1-H2, H1-A1, H1-A2, and H1-A3).

3.2.1 H1-H2

The variable selection process for the H1-H2 model retained the main effects for Sex, Tone, and IP Position, as well as an interaction effect between Tone and IP Position. This means that the model’s explanatory power is reduced if any of these variables is removed, even accounting for the increased parsimony of a smaller model (as the AIC attempts to do). Nonetheless, the only effects whose 95% highest posterior density intervals (HDI) did not include zero in this model were those for Sex=Male, Tone=5,
and the interaction effect Tone=5 × IP Position. The baseline category, to which all comparisons are made, is Sex=Female, Tone=1, IP Position=0, and Region=Rural (this is true for all subsequent models). This does not affect the mathematics of the model, merely the way in which the results are reported.

Credible values of the effect for Sex=Male range from -3.4477 to -0.3326, meaning that male speakers had smaller H1-H2 values than female speakers. The effects for Tone=5 and IP Position can be interpreted as saying that at the beginning of an intonational phrase, neutral tone segments are estimated to have lower H1-H2 (creakier) than the tone 1 baseline, but at the end of an intonational phrase, neutral tone segments are estimated to have higher H1-H2 values (breathier) than those of the baseline.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>estimate</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.252</td>
<td>0.723</td>
<td>-0.810</td>
<td>0.216</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-3.448</td>
<td>-0.333</td>
<td>-1.881</td>
<td>0.009</td>
</tr>
<tr>
<td>Tone2</td>
<td>-0.934</td>
<td>0.219</td>
<td>-0.367</td>
<td>0.215</td>
</tr>
<tr>
<td>Tone3</td>
<td>-0.836</td>
<td>0.256</td>
<td>-0.288</td>
<td>0.301</td>
</tr>
<tr>
<td>Tone4</td>
<td>-0.382</td>
<td>0.631</td>
<td>0.149</td>
<td>0.567</td>
</tr>
<tr>
<td>Tone5</td>
<td>-1.664</td>
<td>-0.198</td>
<td>-0.938</td>
<td>0.013</td>
</tr>
<tr>
<td>‘IP Position’</td>
<td>-0.574</td>
<td>0.724</td>
<td>0.072</td>
<td>0.827</td>
</tr>
<tr>
<td>Tone2:‘IP Position’</td>
<td>-1.188</td>
<td>0.651</td>
<td>-0.237</td>
<td>0.613</td>
</tr>
<tr>
<td>Tone3:‘IP Position’</td>
<td>-1.344</td>
<td>0.472</td>
<td>-0.406</td>
<td>0.384</td>
</tr>
<tr>
<td>Tone4:‘IP Position’</td>
<td>-1.033</td>
<td>0.589</td>
<td>-0.232</td>
<td>0.574</td>
</tr>
<tr>
<td>Tone5:‘IP Position’</td>
<td>0.276</td>
<td>2.388</td>
<td>1.367</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table 3.7: Highest posterior density interval (HDI; 95% with bounds in columns ‘lower’ and ‘upper’) and restricted maximum likelihood parameter estimates for final model of H1-H2

3.2.2 H1-A1

Table 3.8 provides the parameter estimates and HDI bounds for the final model of H1-A1, which included all possible fixed effects. The effects for Tone=2, Tone=3, Tone=4,
Tone=5, IP Position, Tone=2 \times \text{IP Position}, \text{Tone}=3 \times \text{IP Position} \text{ and } \text{Tone}=5 \times \text{IP Position} \text{ have HDI’s that exclude zero and thus merit our special attention. Tone 2 (rising), tone 3 (low falling or low fall-rise), tone 4 as well as neutral tone, have lower H1-A1 at the beginning of the intonational phrase than the intercept, which is tone 1 (high level). For neutral tone, we have reason to believe that H1-A1 increases markedly, anywhere from 2 to 5 dB, as the syllable appears later in the course of an intonational phrase. A similar interaction between tonal category and IP position is found for tone 2 and tone 3, but the effect is not as great (between 0.05 and 2.43 dB per syllable). In addition, the main effect for IP Position indicates that H1-A1 tends to decrease over the course of an intonational phrase.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>estimate</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.715</td>
<td>1.691</td>
<td>-1.051</td>
<td>0.388</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-4.168</td>
<td>1.290</td>
<td>-1.422</td>
<td>0.276</td>
</tr>
<tr>
<td>RegionUrban</td>
<td>-3.194</td>
<td>2.287</td>
<td>-0.412</td>
<td>0.752</td>
</tr>
<tr>
<td>Tone2</td>
<td>-1.573</td>
<td>-0.168</td>
<td>-0.881</td>
<td>0.016</td>
</tr>
<tr>
<td>Tone3</td>
<td>-1.942</td>
<td>-0.609</td>
<td>-1.266</td>
<td>0.000</td>
</tr>
<tr>
<td>Tone4</td>
<td>-1.275</td>
<td>-0.024</td>
<td>-0.629</td>
<td>0.050</td>
</tr>
<tr>
<td>Tone5</td>
<td>-2.980</td>
<td>-1.170</td>
<td>-2.081</td>
<td>0.000</td>
</tr>
<tr>
<td>IP Position</td>
<td>-1.839</td>
<td>-0.251</td>
<td>-1.052</td>
<td>0.009</td>
</tr>
<tr>
<td>Tone2:IP Position</td>
<td>0.198</td>
<td>2.427</td>
<td>1.305</td>
<td>0.023</td>
</tr>
<tr>
<td>Tone3:IP Position</td>
<td>0.059</td>
<td>2.287</td>
<td>1.170</td>
<td>0.040</td>
</tr>
<tr>
<td>Tone4:IP Position</td>
<td>-0.408</td>
<td>1.576</td>
<td>0.577</td>
<td>0.255</td>
</tr>
<tr>
<td>Tone5:IP Position</td>
<td>2.075</td>
<td>4.656</td>
<td>3.390</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 3.8: 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A1

3.2.3 H1-A2

Much as for H1-A1, the final model for H1-A2 included all the fixed effects from the “full” model. Parameter estimates and HDI bounds are provided in table 3.9. The HDI’s for Tone=3, Tone=4, IP Position, Tone=2 \times \text{IP Position}, \text{Tone}=4 \times \text{IP Position}
and Tone=5 × IP Position exclude zero. The main effects for Tone=3 and Tone=4 indicates that tone 3 and tone 4 syllables have lower H1-A2 than the baseline. The main effect for IP Position can be interpreted as saying that H1-A2 values decrease over the course of an intonational phrase, while the interaction effect for Tone=2, Tone=4 and Tone=5 with IP Position means that syllables of these tonal categories are more likely to have higher H1-A2 values as a segment lies later in the intonational phrase.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>estimate</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>10.178</td>
<td>18.876</td>
<td>14.301</td>
<td>0.000</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-4.296</td>
<td>4.663</td>
<td>0.012</td>
<td>0.996</td>
</tr>
<tr>
<td>RegionUrban</td>
<td>-2.645</td>
<td>6.402</td>
<td>1.812</td>
<td>0.458</td>
</tr>
<tr>
<td>Tone2</td>
<td>-1.659</td>
<td>0.016</td>
<td>-0.815</td>
<td>0.058</td>
</tr>
<tr>
<td>Tone3</td>
<td>-1.741</td>
<td>-0.170</td>
<td>-0.954</td>
<td>0.018</td>
</tr>
<tr>
<td>Tone4</td>
<td>-1.527</td>
<td>-0.057</td>
<td>-0.736</td>
<td>0.051</td>
</tr>
<tr>
<td>Tone5</td>
<td>-1.766</td>
<td>0.350</td>
<td>-0.696</td>
<td>0.201</td>
</tr>
<tr>
<td>‘IP Position’</td>
<td>-3.480</td>
<td>-1.632</td>
<td>-2.557</td>
<td>0.000</td>
</tr>
<tr>
<td>Tone2:‘IP Position’</td>
<td>0.247</td>
<td>2.847</td>
<td>1.549</td>
<td>0.020</td>
</tr>
<tr>
<td>Tone3:‘IP Position’</td>
<td>-0.409</td>
<td>2.169</td>
<td>0.897</td>
<td>0.175</td>
</tr>
<tr>
<td>Tone4:‘IP Position’</td>
<td>0.246</td>
<td>2.538</td>
<td>1.366</td>
<td>0.020</td>
</tr>
<tr>
<td>Tone5:‘IP Position’</td>
<td>1.497</td>
<td>4.472</td>
<td>3.043</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 3.9: 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A2

3.2.4 H1-A3

The full model was also selected for H1-A3, as shown in table 3.10. Only two effects have HDI’s that exclude zero, however: IP Position and Tone=5 × IP Position. Taken together, these effects indicate that H1-A3 values decrease across the course of the intonational phrase, except in the case of neutral tone syllables, for which they increase across the IP.
<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>estimate</th>
<th>p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>17.855</td>
<td>23.866</td>
<td>20.805</td>
<td>0.000</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-3.898</td>
<td>2.730</td>
<td>-0.542</td>
<td>0.748</td>
</tr>
<tr>
<td>RegionUrban</td>
<td>-1.253</td>
<td>5.307</td>
<td>2.004</td>
<td>0.234</td>
</tr>
<tr>
<td>Tone2</td>
<td>-0.857</td>
<td>0.513</td>
<td>-0.172</td>
<td>0.625</td>
</tr>
<tr>
<td>Tone3</td>
<td>-1.208</td>
<td>0.083</td>
<td>-0.551</td>
<td>0.095</td>
</tr>
<tr>
<td>Tone4</td>
<td>-0.508</td>
<td>0.699</td>
<td>0.094</td>
<td>0.762</td>
</tr>
<tr>
<td>Tone5</td>
<td>-1.461</td>
<td>0.279</td>
<td>-0.587</td>
<td>0.189</td>
</tr>
<tr>
<td>‘IP Position’</td>
<td>-1.798</td>
<td>-0.271</td>
<td>-1.058</td>
<td>0.007</td>
</tr>
<tr>
<td>Tone2:‘IP Position’</td>
<td>-1.067</td>
<td>1.094</td>
<td>0.008</td>
<td>0.989</td>
</tr>
<tr>
<td>Tone3:‘IP Position’</td>
<td>-1.448</td>
<td>0.706</td>
<td>-0.351</td>
<td>0.522</td>
</tr>
<tr>
<td>Tone4:‘IP Position’</td>
<td>-1.265</td>
<td>0.632</td>
<td>-0.317</td>
<td>0.514</td>
</tr>
<tr>
<td>Tone5:‘IP Position’</td>
<td>0.272</td>
<td>2.758</td>
<td>1.537</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Table 3.10: 95% HDI and restricted maximum likelihood parameter estimates for final model of H1-A3

### 3.3 Discussion

The different models we arrived at above share many points in common, along with some differences. Table 3.11 compiles the signs of the various credible effects, which simplifies the task of comparing the models of different measures to one another. Negative signs indicate lower values on spectral tilt measures, leaning toward creak on the creaky-modal-breathy continuum. Positive signs indicate trends toward breathy voice. Among the social factors, only the Sex effect of a lower mean H1-H2 in men was credible. Recall that H1-H2 has a natural bias toward showing just this effect, due to interactions between fundamental frequency differences across sexes and lack of sex differences in nasal cavity architecture (Slifka 2006). This, combined with the fact that for no other measurement was the estimated effect of being male ever believably divergent from zero, deny us any basis for claiming a difference in creaky voice
production conditioned by gender. Similarly, there was never a believable difference between rural and urban speakers on any measurement.

Among the linguistic factors, second (rising) tones had lower H1-A1 than first (high level) tones. Third (low falling/fall-rise) tones and fourth (high falling) tones had lower H1-A1 and H1-A2 than first tones. Neutral tone syllables had lower H1-H2 and H1-A1 than first tone syllables—at the beginnings of intonational phrases at least. For all measurements—H1-H2, H1-A1, H1-A2, and H1-A3—neutral tone syllables at the end of intonational phrases had higher values, toward the breathy end of the spectrum.

Based on previous literature, we expected 3rd tone and possibly 4th tone to condition lower spectral tilt. We found that 3rd tone, 4th tone, and possibly 2nd tone are associated with lower spectral tilt, as is neutral ‘tone’—except at the ends of intonational phrases. Phonologically, 1st, 2nd and 4th tones are often described as [–Upper] register ([Yip 1980 2002]), so it is unlikely, based on this evidence, that voice
quality is encoding register distinction (as Duanmu (2007) posits). It is important to note that creaky voice doesn’t appear to be related to pitch, either. No spectral tilt measures are meaningfully correlated with actual F0, as evidenced by figure 3.7.

The effect of position in utterance—that spectral tilt measures tend to decrease across the course of an intonational phrase—was as predicted and lines up with research on domain-final phonation in numerous languages, including Mandarin. The most consistent effect across the regressions was that neutral tone syllables were breathier at the end of the intonational phrase than at the beginning, and compared to syllables of other tones.

Finally, the effects involving neutral tone. Why are IP-medial neutral tone syllables creakier? And IP-final neutral tone syllables less creaky, even breathier? Chao (1965) called sentence-final neutral tone particles “intonation carriers” because it was only in their presence, he hypothesized, that pragmatically motivated IP-final perturbations in pitch could be perceived. More recently, Yuan (2004) has proposed models of Mandarin intonation that incorporates pitch targets specified by both tone at the lexical level and intonation at the phrasal level. If neutral tone syllables do not specify any tone targets at the lexical level, perhaps there is functional pressure for them to realize whatever IP-final pitch perturbations are specified at the intonational level. Creaky voice may make it difficult to produce and perceive such fluctuations, since it is often implemented as aperiodic noise. This is, of course, only one possible account. Chen and Xu’s (2006) more recent work argues that neutral tone does specify a (weakly-implemented) mid pitch target. We might derive a prediction from Chen and Xu’s account that the creaky voice of neutral tone syllables in non-final contexts derives from the weakness of the implementation rather than the level of the pitch target. We might further posit, since breathy voice is closer to “open” on Gordon and Ladefoged’s (2001) continuum and the glottis is generally open when at rest, that the
Figure 3.7: F0 plotted against spectral tilt measures, with smoothed regression lines.
weakness of implementation of neutral tones might surface as breathy voice in some IP-final contexts.

3.4 Conclusion

This chapter set out to specify the linguistic and social constraints on creaky voice quality in Beijing Mandarin. The data presented here make a strong case for an effect of IP position and lexical tone on voice quality, with only weak indications of an effect of speaker sex for a social “constraint.” Explaining the divergence in effects between different spectral tilt measurements is key to understanding how best to interpret these findings.

Further work could also attempt to replicate these findings on a different or expanded speaker sample, and also explore what intersection of tone, intonational phonology, and other phonetic/linguistic factors leads to the effects described above. The methodologies I use in this chapter have made analysis of larger amounts of data more economical, the implications of which I will explore in the final chapter.
Chapter 4

Perception and the acoustics of creaky voice

This chapter expands on my study of the social and linguistic distribution of creaky voice from chapter 3, augmenting it with my own ratings of which syllables are creaky and which not. Below, I ask: do the measures of spectral tilt explored in chapter 3 relate concretely to listeners’ judgments of what sounds “creaky”? What does auditory coding of voice quality tell us about the social distribution of creak, and about individual speakers in the corpus?

The perception of creaky voice is very complicated, and it is very difficult, if not impossible, to achieve high degrees of intersubjective agreement as to what counts as creaky or not. This chapter looks at one perceiver’s (my) judgments of what sounds creaky or not. After establishing that my judgments are reasonably consistent with themselves (a step many previous phoneticians have either not conducted or not reported), I look at their relationship to the measures of spectral tilt from chapter 3 and examine some of the variability in this relationship, as well as what it might tell us about the social distribution of creaky voice, as well as the linguistic practices of individual speakers in the corpus.

4.1 Background

Vocal pathologists, who often use percepts of creaky voice, alongside other symptoms, to diagnose voice disorders, have a particular interest in making reliable judgments as
to when a voice sounds creaky or not. Early interest in description of the voice tended to use large lists of qualitative descriptors of a voice’s perceptual quality. These lists tended to be “exhaustive rather than efficient,” as Kreiman et al. (2005, 346) put it. Efforts to provide an empirical basis for perceptual descriptors and purely perceptual descriptions of voice, for example using factor analysis or multidimensional scaling (MDS) in order to study large numbers of descriptors or vocal similarity ratings at once, have largely run aground. In Kreiman et al.’s assessment of a study using a large number of raters, they concluded that “listeners lacked a common notion of what constitutes similarity with respect to voice quality” (352).

Approaches, such as Laver’s (1980), that tried to recenter voice quality analysis on the description of its articulatory basis, rather than its perceptual properties, suffer from similar problems with reliability, with Kreiman et al. reporting that even a three-day workshop in “vocal profile analysis” failed to produce raters able to achieve appreciable inter-rater agreement. Henton and Bladon (1988) both rated their corpus auditorily (while also examining spectrographs for telltale signs of creak), but instead of a formal analysis of disagreement, they produced a consensus rating based on a final, joint listen-through—they do however, state confidently that they encountered “no special difficulties” (16) arriving at the consensus rating. Though auditory coding is commonplace in the sociophonetic study of voice quality (Henton & Bladon 1988; Podesva 2007; Yuasa 2010), given the cautions offered by the likes of Kreiman and others, as well as the lack of explicit quantitative analyses of inter-rater reliability, we should be cautious in applying hand-coding regimes to our data.

Intra-rater reliability in perceptual analysis of the voice has proved somewhat easier to achieve (Kreiman & Gerratt 1996), as it appears that the individual listeners often have comparatively consistent internal models for perceiving voice quality (Kreiman et al. 1992, 2007).
Acoustic approaches to exploring voice quality have been widely applied to cases where languages have voice quality in their phonological systems. Phonologically distinctive voice qualities fall into only a few categories cross-linguistically: modal, creaky, harsh, breathy, lax, epiglottalized, and faucalized (DiCanio 2009, 166). Reducing voice quality variation to this set of labels somewhat sidesteps the problem of qualitatively identifying the voice quality of a certain form. It does not entirely evade controversy, as distinctions of tone, register and voicing are often commingled and difficult to decide upon in a principled fashion (Cao & Maddieson 1992).

Nonetheless, acoustic measures of spectral tilt have had moderate success in describing these categorial differences (Cao & Maddieson 1992, DiCanio 2009), and have also seen some use in sociophonetic explorations of variation in voice quality (Gordon & Ladefoged 2001, Starr & Greene 2006). Still, it is an open question whether the measures of spectral tilt used in chapter 3 have any quantitative relationship with what I auditorily coded as “creaky.” Using the data from the 15 sociolinguistic interviews with young speakers of Beijing Mandarin analyzed in chapter 3, this chapter will look at the relationship between spectral tilt and judgments of creakiness.

4.1.1 Summary of results from chapter 3

We found in chapter 3 that various measures of spectral tilt showed influence from the linguistic environment and possibly the social characteristics of speakers. H1-H2 was lower (creakier) than the baseline for men and on neutral tone in non-IP final contexts. Neutral tone in IP-final contexts had higher H1-H2 (was breathier) than baseline. Results for H1-A1 indicated that tones 2-4, as well as neutral tone, were creakier than baseline, as were IP-final contexts; IP-final neutral tone, tone 2, tone 3 and tone 4 were breathier. H1-A2 indicated that tone 3, tone 4 and IP-final contexts were creakier, while IP-final neutral tone was, again, breathier, as were IP-final tone
2 and tone 4. Finally, the results for H1-A3 showed an effect for IP position (lower H1-A3, creakier), and for IP-final neutral tone (higher value, breathier).

4.2 Methods

I hand-coded the 7164 syllables in the corpus for voice quality, making a forced choice between values of ‘Creaky’ and ‘Not creaky’ after listening to each syllable in isolation. I heard each syllable in random order across the whole corpus, so all speakers were coded at the same time, in order to attenuate the effect of becoming habituated to one speaker’s voice quality.

For the phases of the analysis where I compare these ratings to measures of spectral tilt, I needed a way to characterize the spectral tilt of a syllable. As we reviewed in chapter 1 and chapter 3, spectral tilt is a proxy measure for various articulatory dimensions of creaky voice. They all share in common that they are the difference in amplitude between two harmonics in a power spectrum. H1-H2 measures the difference between the amplitudes of the lowest two harmonics. H1-A1 measures the difference between the amplitude of the harmonic nearest the first formant and the first harmonic. H1-A2 and H1-A3 are the differences between the amplitude of the first harmonic and amplitudes of harmonics nearest the second and and third formants, respectively.

The measurements in chapter 3 were made with vocalic segments as the unit of analysis, rather than syllables (figure 3.3). Many studies have found that the dynamics of phonation across segments and syllables are significant, with some distinctive voice qualities occurring only at the ends or beginnings of the relevant prosodic units (Starr & Greene 2006; Gordon 2001). For this chapter, then, I took the mean of spectral tilt measures in each third of the syllable rhyme, so that each of my hand ratings can
be investigated in relation to spectral tilt at the beginning, middle, and end of the syllable rhyme. These are used as the features (independent variables) the classifiers learned from, with my ratings of creaky/not creaky as the group labels.

After coding the entire corpus, I waited a day and recoded a small subset (n=609) in order to check the consistency and reliability of my judgments. In section §4.3 I will assess the degree to which I agreed with myself on that subset, and go through a confusion analysis to show where inconsistencies in coding were most likely to occur.

Once I establish the intra-rater reliability of my coding, I move on to an analysis of the relationship of those ratings to acoustic measures of spectral tilt, asking whether spectral tilt measures correlate with my determinations of creakiness. I then go a step further and use classificational methods methods to ask if spectral tilt measures can be used to predict whether an observation will be categorized as creaky or not creaky. I attempt two classificational methods in this section: linear discriminant analysis and random forests.

Linear discriminant analysis (LDA) finds a linear combination of the independent variables so as to provide maximally accurate assignment to the groups specified in a grouping variable (Hastie et al. 2009, 106-111). It has been used in the study of vowels in sociophonetics (van der Harst 2011). It is rather simple to implement, and although it is not well-suited for many classification problems, van der Harst (2011) has compared LDA analyses that attempt to predict vowel identity across different Dutch varieties in different contexts: from single-point formant measures and from dynamic (multiple-point) formant measures. Higher correct classification rates from the dynamic measures led van der Harst to recommend dynamic measurements for studies of vowel quality, even for vowels perceived and previously described as monophthongs.
Another predictive classification technique, random forests (Breiman 2001), has become popular recently. Random forests are an extension of tree-based methods for classification and regression (Hastie et al. 2009, 305-317). Trees make a prediction by splitting a set of features (independent variables) into discrete regions and estimating the prevalence of the response in each region. The tree, a binary-branching structure where every node partitions the feature space by some means, for instance by selecting a particular value of one of the independent variables, helps us navigate the space to reach a particular region; at the terminal nodes of the tree, we stop dividing the feature space and make a model of the response. For classification trees, this model is a prediction of the value of the response in that region (for example, whether observations with values in this region are creaky or not creaky).

Like many classification techniques, classification trees can “overfit,” that is, fail to generalize well to new data. Among other advantages, random forests help us avoid overfitting. They do so by “growing” numerous (usually hundreds) of trees on subsamples of the data. Each one of these trees will then classify the observations in the data, assigning it one of your category labels (here, creaky and non-creaky). The label that receives the most “votes”—is selected by the greatest number of trees in the forest—“wins” and is selected as the output of the classifier. As it grows each successive tree, Breiman’s random forest algorithm reserves a third of the sample, keeping it “out of the bag,” and runs the classifier on this data to estimate the accuracy of the tree. The average error rate of each tree on these out-of-bag observations, across the whole forest, is the “out-of-bag error rate,” which is a good unbiased estimate of the classifier’s error rate on unseen data. Tagliamonte & Baayen (2012) have used random forests to develop models and select significant factors affecting variation between was and were in English existential constructions.

98
After reporting the classification analyses, I return to some of the questions of sociolinguistic interest posed in chapter 3 specifically: what are the social and linguistic constraints on the use of creaky voice? Is its occurrence conditioned by speaker sex, region of origin, syllable tone or position in the intonational phrase? The answers are slightly different when we use discrete voice quality rating labels instead of spectral tilt measures, as we will see below.

To demonstrate this I report a mixed-effects logistic regression, with a model selected by a step-down procedure using the Akaike Information Criterion (AIC), akin to the one followed in chapter 3 with voice quality label as the response, and Tone, IP Position (as % of IP), Gender and Region as fixed effects, along with interactions between Tone and IP position and between Gender and Region. Mixed models are becoming more widely used in sociolinguistics because of their ability to account for imbalances introduced by using different amounts of speech from different speakers as well as from words with different underlying frequencies in the language (Baayen 2008; Johnson 2009). In addition, mixed-effects logistic regression (where the response is binary) can increase the accuracy of effect estimates, which may be underestimated by traditional fixed-effects-only models.

The lme4 package in R (R Development Core Team 2009) was used to specify and fit models. lme4 does not implement Monte Carlo Markov Chain estimation for generalized linear mixed models, so it is hard to compute credible intervals akin to those we saw in chapter 3. The package does, however, supply ‘traditional’ z-statistics, which can be used to compute p-values. There are several theoretical issues with relying heavily on these test statistics, chief among them difficulties with estimating the degrees of freedom (see the previous chapter) and a strong assumption that the predictors are distributed normally (No author 2013), but they give a good enough indication of which effects may not be due to chance.
In addition to “replicating” my model of voice quality from chapter 3 (with certain small adjustments to accommodate the different quality of the response), the chapter also offers a cursory exploration of different ways to describe the characteristic voice qualities of individual speakers. Different speakers’ voice qualities “stick out” according to different measures, and in the last stage of analysis in this chapter I examine some such “extreme” speakers, mostly as a way to generate further hypotheses about the social distribution and linguistic description of stylistically meaningful voice quality.

4.3 Results

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Gender</th>
<th>Creaky</th>
<th>Other</th>
<th>DQ</th>
<th>Percent Creak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yijia</td>
<td>Female</td>
<td>6</td>
<td>161</td>
<td>0</td>
<td>3.59</td>
</tr>
<tr>
<td>Yulin</td>
<td>Female</td>
<td>23</td>
<td>392</td>
<td>2</td>
<td>5.54</td>
</tr>
<tr>
<td>Qingwen</td>
<td>Female</td>
<td>31</td>
<td>469</td>
<td>2</td>
<td>6.20</td>
</tr>
<tr>
<td>Haiyu</td>
<td>Female</td>
<td>22</td>
<td>318</td>
<td>1</td>
<td>6.47</td>
</tr>
<tr>
<td>Donge</td>
<td>Female</td>
<td>44</td>
<td>582</td>
<td>10</td>
<td>7.03</td>
</tr>
<tr>
<td>Yueqiong</td>
<td>Female</td>
<td>19</td>
<td>162</td>
<td>4</td>
<td>10.50</td>
</tr>
<tr>
<td>Meiling</td>
<td>Female</td>
<td>87</td>
<td>466</td>
<td>7</td>
<td>15.73</td>
</tr>
<tr>
<td>Xinjuan</td>
<td>Female</td>
<td>105</td>
<td>407</td>
<td>5</td>
<td>20.51</td>
</tr>
<tr>
<td>Luokun</td>
<td>Male</td>
<td>39</td>
<td>349</td>
<td>2</td>
<td>10.05</td>
</tr>
<tr>
<td>Wenhan</td>
<td>Male</td>
<td>84</td>
<td>472</td>
<td>4</td>
<td>15.11</td>
</tr>
<tr>
<td>Yixin</td>
<td>Male</td>
<td>47</td>
<td>263</td>
<td>3</td>
<td>15.16</td>
</tr>
<tr>
<td>Ruiping</td>
<td>Male</td>
<td>114</td>
<td>593</td>
<td>1</td>
<td>16.12</td>
</tr>
<tr>
<td>Weihong</td>
<td>Male</td>
<td>114</td>
<td>524</td>
<td>1</td>
<td>17.87</td>
</tr>
<tr>
<td>Feiyang</td>
<td>Male</td>
<td>93</td>
<td>368</td>
<td>3</td>
<td>20.17</td>
</tr>
<tr>
<td>Yongxiang</td>
<td>Male</td>
<td>97</td>
<td>229</td>
<td>1</td>
<td>29.75</td>
</tr>
<tr>
<td>Quanfeng</td>
<td>Male</td>
<td>132</td>
<td>302</td>
<td>4</td>
<td>30.41</td>
</tr>
</tbody>
</table>

Table 4.1: Creaky/Other ratings per speaker

Table 4.1 shows per-speaker results for syllables rated auditorily creaky, syllables rated not creaky, and syllables disqualified for some reason (too quiet, two voices speaking at once, or too short, for instance). All speakers were creaky at least some
of the time; Yijia the least at 3.593% and Quanfeng the most at 30.41%. Figure 4.1 gives a graphical representation of the same data, showing that my judgments grouped speakers into roughly three clusters—with Quanfeng and Yongxiang together near 30% creak, a group of seven speakers near the middle of the range between 15% and 21% creak, and the remainder of the sample near the bottom, between 3% and 11% creak.

<table>
<thead>
<tr>
<th>Tone</th>
<th>Creaky</th>
<th>Other</th>
<th>DQ</th>
<th>Percent Creak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>131</td>
<td>1183</td>
<td>9</td>
<td>9.97</td>
</tr>
<tr>
<td>2</td>
<td>142</td>
<td>1121</td>
<td>11</td>
<td>11.24</td>
</tr>
<tr>
<td>3</td>
<td>320</td>
<td>1056</td>
<td>12</td>
<td>23.26</td>
</tr>
<tr>
<td>4</td>
<td>225</td>
<td>1984</td>
<td>13</td>
<td>10.19</td>
</tr>
<tr>
<td>5</td>
<td>239</td>
<td>713</td>
<td>5</td>
<td>25.11</td>
</tr>
</tbody>
</table>

Table 4.2: Creaky/Other ratings by tone

<table>
<thead>
<tr>
<th>Syllable Position</th>
<th>Creaky</th>
<th>Other</th>
<th>DQ</th>
<th>Percent Creak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-IP-Final</td>
<td>915</td>
<td>5727</td>
<td>47</td>
<td>13.78</td>
</tr>
<tr>
<td>IP-Final</td>
<td>142</td>
<td>330</td>
<td>3</td>
<td>30.08</td>
</tr>
</tbody>
</table>

Table 4.3: Creaky/Other ratings by IP position (‘IP-Final’=syllable at greater than 90% of IP)

<table>
<thead>
<tr>
<th>Syllable Position</th>
<th>Tone</th>
<th>Creaky</th>
<th>Other</th>
<th>DQ</th>
<th>Percent Creak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-IP-Final</td>
<td>1</td>
<td>119</td>
<td>1123</td>
<td>9</td>
<td>9.58</td>
</tr>
<tr>
<td>Non-IP-Final</td>
<td>2</td>
<td>128</td>
<td>1068</td>
<td>11</td>
<td>10.70</td>
</tr>
<tr>
<td>Non-IP-Final</td>
<td>3</td>
<td>291</td>
<td>1023</td>
<td>11</td>
<td>22.15</td>
</tr>
<tr>
<td>Non-IP-Final</td>
<td>4</td>
<td>199</td>
<td>1886</td>
<td>13</td>
<td>9.54</td>
</tr>
<tr>
<td>Non-IP-Final</td>
<td>5</td>
<td>178</td>
<td>627</td>
<td>3</td>
<td>22.11</td>
</tr>
<tr>
<td>IP-Final</td>
<td>1</td>
<td>12</td>
<td>60</td>
<td>0</td>
<td>16.67</td>
</tr>
<tr>
<td>IP-Final</td>
<td>2</td>
<td>14</td>
<td>53</td>
<td>0</td>
<td>20.90</td>
</tr>
<tr>
<td>IP-Final</td>
<td>3</td>
<td>29</td>
<td>33</td>
<td>1</td>
<td>46.77</td>
</tr>
<tr>
<td>IP-Final</td>
<td>4</td>
<td>26</td>
<td>98</td>
<td>0</td>
<td>20.97</td>
</tr>
<tr>
<td>IP-Final</td>
<td>5</td>
<td>61</td>
<td>86</td>
<td>2</td>
<td>41.50</td>
</tr>
</tbody>
</table>

Table 4.4: Creaky/Other ratings by tone and IP position
Figure 4.1: Individual speakers arrayed on the vertical axis by % syllables rated as creaky. Italic face indicates male speakers, bold face female speakers. Positions are “jittered” (moved randomly) on horizontal axis for better visibility.
Table 4.2, table 4.3, and table 4.4 summarize the ratings for each tone category, according to IP position, and also when both factors are crossed. The results for tone are strikingly similar to those obtained in chapter 3 (see table 3.11). Both tone 3 and ‘tone 5’ (neutral tone) appear to have far higher levels of creak than other tonal categories. As regards IP position, IP-final syllables are nearly three times more likely to appear with creak than other syllables in this corpus.

4.3.1 Intra-rater agreement

<table>
<thead>
<tr>
<th></th>
<th>Crk (recode)</th>
<th>Non-crk (recode)</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crk (orig)</td>
<td>92</td>
<td>9</td>
<td>0.09</td>
</tr>
<tr>
<td>Non-crk (orig)</td>
<td>63</td>
<td>439</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 4.5: Confusion matrix, original coding (by row) vs. recoding (by column; % ‘error’ is rate at which recode does not match original code)

The subsample that I recoded had 88.0597% agreement with the original coding set (Cohen’s $\kappa = 0.6472$). Table 4.5 gives a confusion matrix that shows which categories were most confused with each other. Interestingly, it appears that it was slightly more common for syllables I initially marked as non-creaky to be recoded as creaky than the other way around, indicating that I may have been slightly more ‘permissive’ in what I counted as creaky when I went in for the second round of coding. In any case, 88.0597% agreement is quite substantial, and the roughly even distribution of error across coding categories is also encouraging.

4.3.2 Classification analysis

Having established a degree of intra-rater reliability, the next question to answer is whether there is any quantitative relationship between my labels and the spectral tilt measures extracted earlier. The mean H1-H2, H1-A1, H1-A2, and H1-A3 at each
third of the rhyme in each label category are given in Table 4.6. This table shows that
categories I labeled as creaky had consistently lower values for spectral tilt measures
at all positions within the syllable rhyme. The consistency of this effect is promising,
but the difference between creaky and non-creaky labels in any one cell is much
smaller, across the board, than the variability within categories. Given this fact, how
good can these measures be at predicting assignments of individual observations to
‘creaky’ or ‘not creaky’ labels?

<table>
<thead>
<tr>
<th>Position</th>
<th>Label</th>
<th>H1-H2</th>
<th>H1-A1</th>
<th>H1-A2</th>
<th>H1-A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Creaky</td>
<td>-2.627 (6.514)</td>
<td>-5.193 (8.669)</td>
<td>10.83 (10.97)</td>
<td>18.06 (8.542)</td>
</tr>
<tr>
<td></td>
<td>Non-creaky</td>
<td>-1.822 (6.942)</td>
<td>-3.038 (8.537)</td>
<td>13.34 (11.27)</td>
<td>20.76 (8.456)</td>
</tr>
<tr>
<td>Middle</td>
<td>Creaky</td>
<td>-2.8 (6.931)</td>
<td>-5.432 (9.003)</td>
<td>9.05 (11.5)</td>
<td>17.6 (8.752)</td>
</tr>
<tr>
<td></td>
<td>Non-creaky</td>
<td>-1.872 (6.891)</td>
<td>-3.568 (8.6)</td>
<td>11.67 (11.33)</td>
<td>20.22 (8.494)</td>
</tr>
<tr>
<td>End</td>
<td>Creaky</td>
<td>-2.223 (6.616)</td>
<td>-3.753 (9.001)</td>
<td>11.89 (11.15)</td>
<td>19.16 (8.346)</td>
</tr>
<tr>
<td></td>
<td>Non-creaky</td>
<td>-1.639 (7.041)</td>
<td>-2.299 (8.626)</td>
<td>14.05 (10.69)</td>
<td>21.16 (8.39)</td>
</tr>
</tbody>
</table>

Table 4.6: Spectral tilt measures for creaky/non-creaky label categories, at each third
of the rhyme. Means, with standard deviations in parentheses.

The classification errors of the linear discriminant analysis on measures of spectral
tilt are given in Table 4.7. The analysis achieves high recall for modal tokens, which
are almost all classified accurately, but abysmal recall for creaky tokens; of which only
one observation is correctly categorized. Although the situation is less extreme for the
random forest classifier (Table 4.9), the error rate for creaky tokens is still extremely
high. LDA/random forest with F0 and formant measures added on top of spectral tilt

---

1The comparison between the original coding and the classification ratings for some
category yields four possible results: true positives, a token of the target category cor-
rectly classified; true negatives, a token not of the target category correctly classified as not
belonging to the target category; false positives, a non-member incorrectly classified as a
member; and false negatives, a member incorrectly classified as a non-member. Recall is
generally understood as True positives/N in target category (how many did you get right?),
while precision is True positives/N of tokens the classifier labels as being in target category
(how many of what you got were right?).
did scantly better than their counterparts with only spectral tilt measures, as shown in table 4.8 and table 4.10.

<table>
<thead>
<tr>
<th>Crk (recode)</th>
<th>Non-crk (recode)</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crk (orig)</td>
<td>1</td>
<td>631</td>
</tr>
<tr>
<td>Non-crk (orig)</td>
<td>0</td>
<td>5594</td>
</tr>
</tbody>
</table>

Table 4.7: Confusion matrix, Linear Discriminant Analysis (spectral tilt measures only)

<table>
<thead>
<tr>
<th>Crk (recode)</th>
<th>Non-crk (recode)</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crk (orig)</td>
<td>9</td>
<td>623</td>
</tr>
<tr>
<td>Non-crk (orig)</td>
<td>4</td>
<td>5590</td>
</tr>
</tbody>
</table>

Table 4.8: Confusion matrix, Linear Discriminant Analysis (spectral tilt measures, F0, and F1-3)

<table>
<thead>
<tr>
<th>Crk (recode)</th>
<th>Non-crk (recode)</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crk (orig)</td>
<td>89</td>
<td>968</td>
</tr>
<tr>
<td>Non-crk (orig)</td>
<td>86</td>
<td>5971</td>
</tr>
</tbody>
</table>

Table 4.9: Confusion matrix, random forest classification (spectral tilt measures only), 500 trees, out-of-bag error estimate: 15%.

It would seem, then, that spectral tilt measurements are not very good at distinguishing tokens I labeled as creaky from those I labeled as not creaky. Both the predictions from the linear discriminant analysis and from the random forest classification yielded recall rates above 85%, but this is achieved by putting almost all the observations into the ‘not creaky’ label, which makes up almost 90% of the data. When the random forest classifier does put observations into the creaky label, it does so with only about 50% precision. Since creaky voice is what we are particularly interested in, this is no good. Apparently, the low difference between means of spectral tilt measures by auditory label (usually less than 2) and their high standard deviation (always above 5) means overlap is extremely high for all measures—and there does not
Table 4.10: Confusion matrix, random forest classification (spectral tilt measures, F0, F1-3, 500 trees, out-of-bag error estimate: 14%.

<table>
<thead>
<tr>
<th>Crk (recode)</th>
<th>Non-crk (recode)</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crk (orig)</td>
<td>159</td>
<td>898</td>
</tr>
<tr>
<td>Non-crk (orig)</td>
<td>98</td>
<td>5959</td>
</tr>
</tbody>
</table>

appear to any way of teasing apart the two categories, even in a high-dimensional feature space, at least not with the relatively rudimentary classification techniques employed here.

4.3.3 MODEL OF SOCIAL AND LINGUISTIC FACTORS

A logistic mixed-effects model of auditory creakiness ratings on social and linguistic constraints is reported in table 4.11. The results show both similarities and differences with the models of acoustic measurements of spectral tilt obtained in chapter 3. The variable selection process eliminated the interaction between Tone and IP Position, leaving behind the Gender × Region interaction and main effects for Tone, IP Position, Gender and Region. The resulting effect estimates, as well as the results of (anti-conservative) significance tests, are reported in table 4.11. Tone 3 (low dipping) and neutral tone (aka tone 5) are both likely to be creakier than the baseline (tone 1 [high level]), and there is a strong effect of IP position with later syllables more likely to be creaky. Although the Gender × Region effect was selected as contributing information to the model (contra any of the models in chapter 3), it is not a strong enough effect to be significant.

The Tone × IP Position effect that appeared so consistently in the spectral tilt models was not selected here, and the summary statistics in table 4.14 appear to affirm
Table 4.11: Effect estimates of logistic regression of voice quality rating on social and linguistic factors; negative estimates favor creaky outcomes. Also given are the standard errors of the parameter estimates; the resulting z- and p-values are known to be anti-conservative, so exercise caution.

|                          | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------------|----------|------------|---------|----------|
| (Intercept)              | 4.00     | 0.28       | 14.23   | 0.00     |
| Gender=Male              | -0.52    | 0.38       | -1.37   | 0.17     |
| Region=Urban             | 0.18     | 0.40       | 0.44    | 0.66     |
| Tone=2                   | -0.12    | 0.16       | -0.74   | 0.46     |
| Tone=3                   | -1.28    | 0.14       | -9.18   | 0.00     |
| Tone=4                   | -0.18    | 0.14       | -1.25   | 0.21     |
| Tone=5                   | -1.03    | 0.17       | -6.14   | 0.00     |
| IP Position              | -1.81    | 0.16       | -11.59  | 0.00     |
| Gender=Male : Region=Urban| -0.81    | 0.55       | -1.49   | 0.14     |

that neutral tone syllables are not differently affected by IP finality than any other tone category. It is unclear if this finding contradicts the results from chapter 3.

It is possible the higher spectral tilts of final tone 5 syllables are due to some other difference in voice quality that cross-cuts the “creaky”/“not creaky” distinction I draw in this chapter. It is also possible that difficulties in measuring spectral tilt for tone 5 syllables, which in all prosodic positions tend to be shorter and quieter, were simply compounded in final position, resulting in numerous spurious readings. For this effect to have held over all 4 measures of spectral tilt and still be spurious, however, would require somewhat deeper explanation, which may have to await further investigation.

4.3.4 Examination of individual speakers

Since I also want to explore variability in voice quality between individuals in my speaker sample, I first did some work to relate the creakiness ratings to measures of spectral tilt on a per-individual basis. This necessitated versions of the spectral
Figure 4.2: Standardized speaker mean H1-H2 plotted (y-axis reversed) against percent “Creaky” ratings; $R^2=0.075, 0.111, 0.091$, from top to bottom
Figure 4.3: Standardized speaker mean H1-A1 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.066, 0.09, 0.09$, from top to bottom
Figure 4.4: Standardized speaker mean H1-A2 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.007, 0.009, 0$, from top to bottom
Figure 4.5: Standardized speaker mean H1-A3 plotted (y-axis reversed) against percent “Creaky” ratings, $R^2=0.004, 0.006, 0.012$, from top to bottom
tilt measures that were scaled so as to be comparable to each other. I divided each measure of spectral tilt across the corpus by its standard deviation, and then subtracted the sample mean, to produce a normalized “z-score” version of the measure. Comparing the means of these normalized scores for each speaker to the hand-coded creakiness ratings (see figures 4.2, 4.3, 4.4 and 4.5), we see that some measurements have a higher correlation with creakiness ratings than others. The correlation between spectral tilt and creakiness ratings is stronger for H1-H2 and H1-A1 than for H1-A2 and H1-A3. This is a logical division, since H1-H2 and H1-A1 measure differences in the lower frequencies of the spectrum, while H1-A2 and H1-A3 reflect longer-range differences in the spectrum, and H1-H2 in particular is usually, though not universally stronger at predicting voice quality distinctions (Kreiman et al. 2007). Three speakers in particular are picked out as extreme on these measures, though they are all only middling in hand-coded ratings of creakiness: Wenhan is a strong outlier on H1-H2 and H1-A1, and Haiyu and Yueqiong are also prominent on H1-A2 and H1-A3.

Adding in the two speakers with near 30% creakiness in hand-coded ratings, this gives us three “extremely” creaky men—Quanfeng, Yongxiang, and Wenhan—and two extremely creaky women—Yueqiong and Haiyu. Interestingly, the three men were all from urban neighborhoods close to Beijing’s core (Xicheng, Haidian—which straddles urban and suburban divides, but Yongxiang was from the solidly urban ‘tech hub’ of Zhongguancun—and Chaoyang), while Haiyu and Yueqiong are from relatively far-flung and rural suburbs (Huairou and Changping, respectively). The women, who are friends and came into the interview together, seemed qualitatively quite creaky to me, so it was surprising to me that they were not rated as “creaky” when I hand-coded the data. Perhaps, though, their preferred non-modal phonation was something subtly different—whatever was picked up by the “mid-range” spectral tilt measures H1-A2 and H1-A3. That this voice quality might be distinct in form and function from
that used by the urban men is circumstantially corroborated by the large apparent differences in lifestyle and economic activity they reported (for instance, Haiyu and Yueqiong worked summer jobs waiting tables while the men did not plan to work during the summer), as well as in other linguistic behavior (the pronunciation of the -ai final seemed to be particularly variable). Though such ethnographic detail and sociophonetic hypotheses deserve fuller attention from a more in-depth investigation, it is worth noting that different measures of spectral tilt might be registering variability that is meaningfully distinct for listeners.

4.4 Discussion and Conclusion

This chapter’s investigation has established a few facts: first, that a rater’s (my) impressions of what counts as creaky voice and what does not are reasonably self-consistent. This may seem a rather low bar to set for oneself—typically we look for patterns that can be intersubjectively verified by members of the speech community under study, or at least by native speakers of the target language! But prior research on the intersubjective reliability of voice quality rating has lowered our hopes enough in this regard that intra-rater reliability is at least a good start. There is some research on individual differences in voice quality perception [Kreiman et al. 1992], but none based in speech communities or communities of practice—a wonderful next step would be to integrate the study of voice quality perception into an ethnographically informed community investigation, to see whether voice quality perceptions (especially in communities where voice quality is sociolinguistically meaningful) are more highly ‘cohesive’ in such collectives [Fruehwald & MacKenzie 2011].

Another fact established here, building on the first, is that my labeling of creaky/not-creaky voice quality is empirically linked to ‘objective’ measures of spec-
tral tilt. Although measures of spectral tilt have proved useful in distinguishing linguistically distinctive voice qualities, and have been employed in sociolinguistic applications, it has been rare for studies of voice quality variation to compare auditory classifications to acoustic measurements. Although the cautions about intersubjective reliability raised above compel caution when describing exactly what construct is being verified by this correlation—we are, at base, modeling my ratings of voice quality—this is an important conclusion, and builds on earlier sociophonetic work that has either used auditory methods (Henton & Bladon 1988; Podesva 2007) or acoustic ones (Teshigawara 2003), though Starr and Greene (2006) did verify their distinction between modal and breathy with measures of spectral tilt.

The step beyond identifying the correlation between category labels and acoustic measures, i.e. to automatic classification based on acoustic data, treads into a difficult problem area, where active research continues to make incremental advances (e.g., Kane et al. 2013 which achieves recall of creak between 60% and 80%—using a classification tree analysis akin to the random forest used here). It is clear, however, that however useful spectral tilt measures are, they are not sufficient to predict rater categorizations on their own.

This is an important point to keep in mind, as automatic annotation of sociolinguistically interesting variables has achieved some degree of success in recent years—as, for instance, in the case of (ING) (Yuan & Liberman 2011). As access to large amounts of data becomes more and more economical, the problem of annotating it will only become more costly if we stick to manual-only methods of annotation. Moving sociophonetics to work with easily-automated acoustic surrogates of variables of interest, as I have done in chapter 3, is one solution, but for many perceptual variables, one-dimensional acoustic correlates either do not exist (for instance, for /t/-dropping, which surfaces as an absence, as a glottal stop, or even as creaky voice)
or are extremely variable (as in the present case). Sociolinguists would do well to stay on the lookout for independently validated methods for annotating audio data with qualitative category labels, as they continue to be devised. This also entails a deeper relationship between sociolinguistics and computational linguistics, one which can be mutually beneficial for research and practical ends in both fields (see, for example, Zheng, Sproat, Gu, Shafran, Zhou, Su, Jurafsky, Starr & Yoon [2005]).

Caution should be exercised, however, especially in small and medium-sized studies, not to get “too far” away from the data, as mismatches between one method of annotation and analysis or another can be uninteresting noise or the result of meaningful difference. Ethnographic involvement with participants, attention to discourse, and other “close” methods can help distinguish these cases. Any of these interventions—exploring where creak appears in discourse and interactional contexts, establishing its particular interactional or informational functions, alongside deeper ethnographic engagement with the participants—would help us a long way down the road toward greater understanding of creak’s meanings and functions.
Chapter 5

Linguistic context and the social evaluation of creaky voice

Having established that there are definite linguistic constraints and possibly social constraints on the distribution of voice qualities in Mandarin Chinese, we now turn to the effect of creaky voice on listeners’ evaluations of speakers. This chapter uses experimental methods to explore what social stereotypes are associated with creaky voice quality in Mandarin Chinese. Most fundamentally, it asks: how does the linguistic context in which variants occur relate to their social meanings?

5.1 Background

The direct antecedents of this chapter lie in the “matched guise” paradigm, which presents listeners with language stimuli (usually audio recordings) and asks them to evaluate traits of the speaker based on the linguistic evidence. Lambert’s (1966) original ‘trick’ was that some of the stimuli were actually produced by the same speaker, but in different languages in which they were equally competent. This technique can be used to uncover “language attitudes,” language users’ sometimes covert evaluations of different language varieties. In more recent work, it has become common to use various forms of digital audio manipulation to achieve a similar—and often more tightly controlled—effect, in order to understand people’s evaluations of variability within a single language.

The matched-guise technique has a long history in sociolinguistics. It was often applied to the gauging of attitudes toward different languages, but was also applied to
dialectal variation within a single language (Giles & Bourhis 1976; Ball 1983; Rickford 1985). In the latter case it has provided data that urged on the development of the influential Communication Accommodation Theory (Giles et al. 1991; Giles & Ogay 2007). One largely consistent finding of the language attitudes paradigm is that attitudes towards language varieties seem to break down on axes of perceived intelligence or capability and of perceived friendliness or social approachability (Campbell-Kibler 2010). These studies focused on attitudes toward entire varieties.

It is also possible to use the matched guise technique to investigate the social meanings of individual variables. Campbell-Kibler’s (2007, 2010) deftly executed investigation of verbal (ING) in English used the matched guise technique (see also Fridland et al. 2004 on perceptions of vowel variation in US English for methods in a similar vein). With audio recordings of sociolinguistic interviews which she conducted, Campbell-Kibler extracted several paragraph-length excerpts of English speakers using verbal (ING). Having found sufficient examples of both “standard” velar ING (-ing) and alveolar (-in’), she used digital splicing to transform -in’ s into -ing’s and vice versa. Forthcoming work from Podesva et al. (n.d.) used the same splicing techniques to investigate evaluations of released /t/ in the speech of well-known U.S. politicians.

One of the most in-depth investigations of the use of creaky voice in US English (Yuasa 2010) includes a survey component that bears special scrutiny for its similarity to my own investigation. Yuasa used audio recordings of a California woman talking about food and gave participants two clips to rate—one mostly creaky and one mostly non-creaky. Raters responded that the woman’s creaky voice sounded educated, casual, genuine, and non-aggressive. In open-ended responses they commented that the creaky clip sounded “professional,” “upwardly mobile,” and “urban.” Yuasa’s work on the social perceptions of creaky voice in the US is a fascinating first pass at
assessing its social meanings, but there are several ways in which its methods could be made tighter and more systematic. Participants heard both creaky and non-creaky speech, were made aware that creaky voice was the object of study, they evidently knew that the speaker was the same in each guise and there was only one stimulus in each voice quality. It is important to preserve the matched guise paradigm’s mechanisms for shielding the object of study from the conscious awareness of participant. Also, to the extent practicable, it is important to use a wider variety of stimuli for listeners to evaluate, to help control for any semantic or pragmatic idiosyncrasies of individual stimuli. This chapter’s methodology aims to implement at least these controls, as described below.

5.2 Methods

The experiment followed a “matched guise” design, in which participants listened to short speech recordings, in either of two guises: creaky, where creaky voice quality had been artificially introduced through splicing, and plain, where ‘modal’ or normal voice quality had been introduced by the same means.

For each stimulus, participants evaluated the speaker using Likert scale assessments on several dimensions, described below. They also gave optional free responses with short descriptions of the speaker, which will not be analyzed in this chapter.

5.2.1 Design

The main independent variable is Guise, a between-subjects variable with creaky and plain conditions. Other independent variables include Position of Modification (IP Position), a between-subjects variable controlling whether a modification is made in IP-medial (at the 4th syllable) or IP-final (at the 10th and last syllable) position, and
Lexical Tone, a within-subjects variable that determines the tone of the syllable to be modified. Lexical Tone compared tone 3, a “low” tone often correlated with creaky phonation (see chapter 3), with tone 4, a high falling tone. Each participant heard a mix of creaky and modal stimuli in random order, with no one encountering the same sentence in both guises. Stimuli were presented one at a time in an online survey, along with orthographic transcription and controls that allowed the participant to replay the stimulus as many times as desired. Questions were displayed below and responses were provided on radio buttons and in text fields.
5.2.2 Stimuli

In order to investigate the effects of guise on ratings, as well as the potential interactions between tonal and prosodic context and these ratings, the design and creation of stimuli had to satisfy several requirements.

One was prosodic consistency. All stimuli were controlled for length, set at ten syllables. IP-medial modifications were made at the fourth syllable, while IP-final modifications happened at the tenth syllable. Although this obviously fails to neutralize all possible sources of prosodic variability—which might include preceding and following tone (although tone sandhi-inducing environments were excluded), duration, prosodic pitch targets, etc.—the distinction between IP-finality and mediality is gross enough for our purposes here.

Another desideratum was consistent natural-seemingness across guises. The stimulus production process sought to produce natural-sounding modified speech for both plain and creaky guises. Both guises were digitally manipulated, so as to reduce the chances that differences in participants’ responses were due to the comparative unnaturalness of a modified guise.

A diversity of tonal contexts (with critical syllables in both tone 3 and tone 4) and segmental contexts was also desirable, so a list of syllables was devised from which stimulus sentences would eventually be constructed. The stimulus production process that seeks to satisfy all these criteria is described below.

The sentences themselves were generated by somewhat novel means using the “crowd-sourcing” platform Amazon Mechanical Turk (see Schnoebelen & Kuperman 2010 for a linguist-friendly introduction). “Turkers” who listed their native language as Chinese were asked to devise sentences ten syllables in length, using a given syllable (one of the above-described set of critical syllables), and deploying the syllable as
either the fourth or the tenth in the utterance. They were asked to make a single phrase, with no commas (phrasal or list commas), semicolons or other punctuation, to encourage users to write sentences easily read under a single IP.

Once these sentences were produced, I examined them for non-compliance with the task requirements as well as obvious syntactic errors or anomalies of meaning. They were then sent back to Mechanical Turk for norming. A survey asked six questions, listed below, which raters responded to using Likert scales scored from 1 to 6 or text fields:

1. Please describe your reaction to the sentence:
   
   (a) I hate it ...I approve of it
   
   (b) I think it is boring ...I think it is very vivid
   
   (c) I think it is fake ...I think it is real

2. Based on this sentence, which description do you think best suits the author?
   
   (a) Not interested in the topic ...Very interested in the topic
   
   (b) Not knowledgeable in the topic ...Knowledgeable in the topic
   
   (c) Not very friendly in general ...Very friendly in general

3. Is there anything strange about this sentence?

   If raters (each candidate sentence was rated at least six times) answered that they found anything strange about a sentence, it was discarded. The quantitative responses were then scored by placing them into two categories: The deviations of $1a$, $1b$, $2a$, and $2c$ from the candidate group mean for each question were averaged and subtracted from 3.5 (the mean of the possible ratings) for an approximation of
the stimulus’ “affective intensity,” while the scores for 1c and 2b were averaged for a measure of the stimulus’ “believability.”

The sum of these scores was taken and the highest-scoring stimulus using a particular target character in a particular position was used. After throwing out two of the 25 candidate syllables, which yielded no usable sentence stimulus candidates, a set of 46 sentences was identified for further development.

Two fluent native speakers of Beijing Mandarin, both female, were asked to produce the raw speech data for each audio stimulus. They recorded each sentence four times, and recorded isolated tone 1 (high-level) versions of each target syllable. For a large number of the sentence-type productions, one production of the target was creakier than another. In such cases, I spliced target syllables into a carrier (a third production, the original carrier of neither spliced target syllable). If the results did not sound obviously artificial to me or a native Mandarin speaking listener, the pair of modified stimuli was included in the experimental materials. Some sentence items produced no usable candidates for one or both speakers. See Appendix B with a table of the stimuli and summary of which carriers had viable creaky-plain pairs.

The experimental conditions included IP Position of the target (whether IP-final, at the tenth syllable or IP-medial, at the fourth) and Guise (whether the modification introduced creaky or non-creaky phonation). The Tone condition controlled whether the syllable manipulated belongs to Mandarin’s low dipping tone (tone 3) or high falling tone (tone 4). The 106 finished stimuli were separated into four lists, numbering either 26 or 27. Lists were populated with a mix of guises, but no one speaker speaks the same sentence more than once in any one list. Participants were randomly assigned to a list and heard the stimuli in randomized order.
<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>说话人很高兴～不高兴</td>
<td>The speaker is happy ~ is unhappy</td>
</tr>
<tr>
<td>说话人有文化～没有文化</td>
<td>The speaker is cultured (i.e. educated) ~ is not cultured</td>
</tr>
<tr>
<td>说话人富裕～贫穷</td>
<td>The speaker is rich ~ is poor</td>
</tr>
<tr>
<td>说话人情绪很高～情绪很低</td>
<td>The speaker is enthusiastic ~ is unenthusiastic</td>
</tr>
<tr>
<td>说话人社会地位很高～很低</td>
<td>The speaker’s social position is high ~ is low</td>
</tr>
<tr>
<td>说话人来自城市～来自农村</td>
<td>The speaker comes from the city ~ a village</td>
</tr>
<tr>
<td>说话人知识渊博～知识浅薄</td>
<td>The speaker’s knowledge is vast ~ is limited</td>
</tr>
<tr>
<td>说话人很感兴趣～不感兴趣</td>
<td>The speaker is interested ~ is uninterested</td>
</tr>
</tbody>
</table>

Table 5.1: Response items (both poles)

5.2.3 Participants

Of 19 participants who started the survey, there were 14 participants in this experiment who provided full responses and will be analyzed here. Participants were recruited online through Facebook, friend networks, and professional contacts. All were native Mandarin speakers with normal hearing. Demographic data were collected on participants’ gender, linguistic background (specific home dialect, exposure to other languages), birthplace and present-day location, as well as level of education. Information about the participants is displayed in tabular format in table B.1 in Appendix B.

5.2.4 Responses

On each stimulus screen, where participants could play and replay the stimulus audio, as well as read it in standard Simplified Chinese orthography, participants also answered a series of nine questions. The response items were developed in consultation with informants in Beijing and in early iterations of the online interface.
Eight of the response items were six-point Likert scales between different statements about the speaker, along with a number of free response items not analyzed in this chapter. See table 5.1 for a list of the questions asked, in Chinese and English.

5.2.5 MODELING AND ANALYSIS

Some creaky-plain stimulus pairs are available only for one speaker or another, and there are not equal numbers of stimuli in each condition of each factor. These imbalances require redress in the methods of analysis. Mixed-effects regression using random intercepts for respondent, speaker, and each creaky-plain stimulus pair help account for the extraneous variability introduced by these factors. One large model of all the responses was run first (the “big model”), followed by regressions for each individual response item (question) in turn.

The variable selection process was a step-down procedure from a full model, ending at a potentially reduced model with a lower number of parameters. The full model in the “big model” of all responses started with two three-way interactions: Tone × Guise × Question and IP Position × Guise × Question, alongside all main effects. For the per-response regressions, fixed effects were Tone, IP Position, Guise, and both two-way interactions Tone × Guise and IP Position × Guise. The step-down procedure proceeded by dropping the parameter resulting in the greatest reduction in AIC and stopping when no reduction in AIC was possible. If the procedure reduced the model to main effects only, with no interactions with guise, the model containing main effects for Tone, IP Position, and Guise was selected for reporting.
5.3 Results

The mean response values and standard deviations for each guise of each stimulus group are given in table 5.2.

<table>
<thead>
<tr>
<th>Question</th>
<th>Creaky</th>
<th>Plain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Charming - Not charming</td>
<td>2.845 (1.099)</td>
<td>2.846 (1.213)</td>
</tr>
<tr>
<td>2 City - Village</td>
<td>3.6 (1.467)</td>
<td>3.713 (1.596)</td>
</tr>
<tr>
<td>3 Educated - Uneducated</td>
<td>2.955 (1.316)</td>
<td>2.882 (1.261)</td>
</tr>
<tr>
<td>4 Enthusiastic - Unenthusiastic</td>
<td>2.955 (1.316)</td>
<td>2.882 (1.261)</td>
</tr>
<tr>
<td>5 High social status - Low social status</td>
<td>2.985 (0.9214)</td>
<td>3.01 (1.055)</td>
</tr>
<tr>
<td>6 Interested - Uninterested</td>
<td>3.18 (1.413)</td>
<td>3.169 (1.424)</td>
</tr>
<tr>
<td>7 Knowledgeable - Not knowledgeable</td>
<td>3.065 (0.9462)</td>
<td>3.072 (1.096)</td>
</tr>
<tr>
<td>8 Rich - Poor</td>
<td>3.07 (0.9642)</td>
<td>3.051 (1.139)</td>
</tr>
</tbody>
</table>

Table 5.2: Means and standard deviations of response values for each guise of each stimulus group. Labels are given in the same left-right order that participants saw them in. Lower values indicate responses tending toward the left-hand label, higher values the right-hand label.

The “big model” mixed-effects regression model for all response values selected according to the criteria in section §5.2 is displayed in table 5.13 at the end of the chapter. It includes a three-way interaction between IP Position, and Guise, and Question (the factor for which evaluation dimension a participant was rating on). Tone was not selected in this regression. In this table, IP-final Creak has a credible negative effect on ratings of ‘Enthusiastic - Unenthusiastic’ (values closer to zero indicate more enthusiastic) compared to the baseline. Individual regression models for responses to each question are reported below in table 5.3 through table 5.11.

The models for ‘Charming’ (table 5.3), ‘Social status’ (table 5.4), ‘Rich’ (table 5.5), and ‘Knowledgeable’ (table 5.6) did not include any interactions with Guise and are thus not very interesting in light of our research questions.

The models of responses to ‘Interested’ (table 5.9) and ‘Enthusiastic’ (table 5.11) included interactions with Guise. For ‘Interested - Not interested,’ the interaction
Table 5.3: Mixed effects model of response values for 'Charming - Not charming'

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.62</td>
<td>5.28</td>
<td>2.88</td>
</tr>
<tr>
<td>tone4</td>
<td>-0.22</td>
<td>0.13</td>
<td>-0.04</td>
</tr>
<tr>
<td>position10</td>
<td>-0.17</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.19</td>
<td>0.13</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

Table 5.4: Mixed effects model of response values for 'High social status - Low social status'

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.16</td>
<td>4.49</td>
<td>2.89</td>
</tr>
<tr>
<td>tone4</td>
<td>-0.19</td>
<td>0.12</td>
<td>-0.03</td>
</tr>
<tr>
<td>position10</td>
<td>-0.04</td>
<td>0.29</td>
<td>0.14</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.08</td>
<td>0.20</td>
<td>0.06</td>
</tr>
</tbody>
</table>

between Guise and IP Position was selected as adding information to the model. The main effects show a tendency for creaky guises to be judged as less interesting and for sentences with final modifications to also be judged as less interesting. The interaction effect specifies that speakers in creaky guises with IP-final modification were rated as being more interested in the topic than the main effects for creaky guise and IP position would indicate together. This relationship can be seen in table 5.10 and figure 5.2, where plain, IP-medial guises and creaky, IP-final guises have the lowest values.

The interaction between Guise and IP Position also held for ‘Enthusiastic - Unenthusiastic,’ with main effects for creaky guise and IP-final modification on their own pushing judgments in the direction of ‘Unenthusiastic’ and an interaction effect showing that stimuli in creaky guises with IP-final modification were judged as more
Figure 5.2: Predicted values for Interested - Not Interested (fixed effects only)
Table 5.5: Mixed effects model of response values for ‘Rich - Poor’

<table>
<thead>
<tr>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.82</td>
<td>6.97</td>
</tr>
<tr>
<td>tone4</td>
<td>-0.12</td>
<td>0.22</td>
</tr>
<tr>
<td>position10</td>
<td>-0.09</td>
<td>0.27</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 5.6: Mixed effects model of response values for ‘Knowledgeable - Not knowledgeable’

<table>
<thead>
<tr>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.26</td>
<td>6.44</td>
</tr>
<tr>
<td>tone4</td>
<td>-0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>position10</td>
<td>-0.07</td>
<td>0.31</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.07</td>
<td>0.21</td>
</tr>
</tbody>
</table>

‘Enthusiastic.’ The effect size (-0.4125) of the interaction effect was substantially similar to the effect for ‘Interested.’ The model’s predicted ratings for all combinations of position and guise are displayed in table 5.12 and figure 5.3.

5.4 Discussion and Conclusion

Interestingly, native Mandarin listeners showed influence from linguistic context, specifically prosodic context, on their evaluations of speakers using creaky voice. These differences in evaluation surfaced along two response dimensions—interest and enthusiasm. Models for these dimensions indicate IP position affects the difference in ratings between creaky and plain guises, with the direction of the relationship between guises reversed for both items. The most interested and enthusiastic stimuli were those
Figure 5.3: Predicted values for Enthusiastic - Unenthusiastic (fixed effects only)
in plain, IP-medial conditions, but the next most enthusiastic and interested stimuli were those in the creaky, IP-final conditions.

It bears mentioning that for none of the response items was a main effect of Guise credibly non-zero (that is, none of the credible intervals for Guise excluded zero). This, combined with the tricky interaction effects for the models that included them, actually leaves us somewhat ill-equipped to answer our first research question about the social meaning of creak. This experiment provides little evidence of what social stereotypes are associated with creaky voice per se in Mandarin. What we have instead are effects on speaker evaluations from linguistic variants in particular contexts. This is consonant with work that emphasizes the intense context-sensitivity of the social meanings of variation (Campbell-Kibler 2007; Podesva et al. n.d.).
<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.14</td>
<td>5.84</td>
<td>2.94</td>
</tr>
<tr>
<td>position10</td>
<td>0.04</td>
<td>0.71</td>
<td>0.37</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.03</td>
<td>0.68</td>
<td>0.33</td>
</tr>
<tr>
<td>position10:is.creakyTRUE</td>
<td>-0.86</td>
<td>-0.02</td>
<td>-0.43</td>
</tr>
</tbody>
</table>

Table 5.9: Mixed effects model of response values for ‘Interested - Uninterested’

<table>
<thead>
<tr>
<th></th>
<th>Plain</th>
<th>Creaky</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP-medial</td>
<td>2.94</td>
<td>3.27</td>
</tr>
<tr>
<td>IP-final</td>
<td>3.31</td>
<td>3.21</td>
</tr>
</tbody>
</table>

Table 5.10: Predicted values for Interested - Not Interested (fixed effects only)

Despite these difficulties, this data helps us address important questions. First, it seems reasonable to conclude that the social meaning of creak in Mandarin has something to do with enthusiasm and interest. Bearing in mind that further work will also have to sort out the relationships between these traits, which in raters’ minds may have latent links with one another, we should still explain why impressions of speakers appear to be affected on these dimensions and no other.

Interest (感兴趣 ‘feels interest’) and enthusiasm (情绪高 ‘emotion elevated’) both denote mental or emotional states of the speaker. The ‘Interested’ item relativizes interest specifically to the speech topic, while the ‘Enthusiastic’ item asks participants to rate the speaker’s state without specific relation to the topic. Still, both items entail variation in the intensity of emotional states. It has been suggested that the low apparent pitch and decreased amplitude of creaky voice can be schematized as signifying “low energy” [Brown & Levinson 1987; Gobl & Ní Chasaide 2003]. However, schemata which frame creak as an expansion of the pitch range or an implementation
Table 5.11: Mixed effects model of response values for ‘Enthusiastic - Unenthusiastic’

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.62</td>
<td>4.83</td>
<td>2.76</td>
</tr>
<tr>
<td>position10</td>
<td>0.06</td>
<td>0.76</td>
<td>0.41</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.21</td>
<td>0.48</td>
<td>0.13</td>
</tr>
<tr>
<td>position10:is.creakyTRUE</td>
<td>-0.82</td>
<td>0.01</td>
<td>-0.41</td>
</tr>
</tbody>
</table>

Table 5.12: Predicted values for Enthusiastic - Unenthusiastic (fixed effects only)

<table>
<thead>
<tr>
<th></th>
<th>Plain</th>
<th>Creaky</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP-medial</td>
<td>2.76</td>
<td>2.89</td>
</tr>
<tr>
<td>IP-final</td>
<td>3.17</td>
<td>2.88</td>
</tr>
</tbody>
</table>

of “super-low” pitch target (Mendoza-Denton 2011) (see also Podesva 2007 and Sicoli 2010) provide a basis for an alternative semiotization—one of increased effort involved in veering further from “baseline’ than normal.

We should be open to the possibility that such alternative schemata for semiotizing creaky voice may be active at the same time, becoming relevant as circumstances, discourse contexts, and linguistic environments change. Autosegmental models of tone and intonation posit pitch targets as phonological entities in their own right and although extant autosegmental models and transcription systems of Mandarin intonation (Peng et al. 2006) have not advanced far enough to confirm the existence of phonologically specified pitch targets at the intonational level, evidence from F0 dynamics at the right boundary of the IP suggests that there may at least be low pitch targets at the ends of most Mandarin intonational phrases.

Regardless of the level and nature of the specification, if IP ends call on a speaker to implement a low pitch, then the schema that semiotizes pitch as effort, rather than
low energy, becomes relevant in a way that it is not in a situation where low pitch is not otherwise specified. Thus the fact that, in IP-medial modifications, normal guises were rated as notably more interested and enthusiastic is not surprising if we infer that raters were working with the creak-as-low energy framework. With IP-final modifications, however, listeners who expect a speaker to be implementing a low pitch may be comparably impressed at the apparent “extra” effort involved in achieving creak as opposed to maintaining modal phonation, and interpret that effort as interest or enthusiasm.

These implications are invigorating, and I expand on them in the next chapter to speak more generally on how they inform our understanding of variation and social meaning. But there are several sources of uncertainty and possible error that limit the scope of these findings. In particular, high confidence would require a higher number of respondents, and limiting recruitment to participants with native Beijing Mandarin listening would eliminate a lot of variability in rater experience that could color their evaluations. It proved impracticable in the online interface to clearly assess participants’ exposure to Beijing and Beijing Mandarin, much less the putonghua spoken in the city and outside areas, which shows considerable variability (in production and exposure) even inside the Beijing municipal area. This means that incorporating listeners from diverse geographical and dialectal backgrounds may have introduced unknown variability into the models and estimates above.

I was surprised that there was not a single response dimension for which a credible main effect of voice quality guise was found. It would help us to know what creak “means” in the general case in order to better understand the contextual interactions at work in these data. Logically speaking it is possible that there is no “general case”—that the functions of creaky voice mutate so radically in different contexts that there is no underlying meaning. While convincing arguments exist for this kind of shiftiness
relative to discourse context (Chun & Podesva 2010; Mendoza-Denton 2011), there is a high burden of proof for establishing this kind of mutability relative to linguistic or prosodic context. We would not necessarily expect linguistic structure in and of itself to provide any kind of ‘partially congruent text-default’ (in the parlance of Agha 2007a), that is, competing meanings that might cancel out or inflect social meanings ordinarily enacted by creaky voice. This is an important point, one that separates the account here from one of enregisterment. When a particular form or “instance” of structure modifies, by dint of its very presence, the overall indexical “effect” of a text, this modification is precisely the form’s text-default meaning. If this effect is durable across contexts (“detachable”), it is likely that the form’s text-default meaning is the historical product of enregisterment. In the case of prosodic context in this chapter, however, structure does not act alone. IP-medial position merely affords a schema for imputing the meaning of creak which is different from the schema afforded by IP-finality. Both position in the IP as well as the presence of creak are necessary to arrive at the final interpretation, and in this way IP-position and creak are parts of a “composite” (Enfield 2012) whose inner workings are more than just additive.

It is in this way that different linguistic and prosodic contexts may favor different competing semiotic schemata, or frameworks for deriving meaning. This bears a striking resemblance to Silverstein’s account of how r-less speech is evaluated in New York’s Lower East Side (reviewed in chapter 2), in which he also proposes that language ideology provides an evaluative rubric against which the meaning of r’s absence can be interpreted as iconic (or rhematic) of lack or absence. This view would be compatible with the increased focus on iconicity in social meanings of language (Irvine 2005; Eckert 2012), since it allows for iconicity to coexist with convention and eschews overly essentialist accounts of meaning. As an example of such naturalistic, essentialist theories, we need look no further than Ohala’s (1983) and Gussenhoven’s
“biological codes” for interpreting pitch and intonation, which flatly ignore the facts of enregisterment and the existence of “arbitrary” social meanings. Although the specifics of a dynamic, schema-based approach to these kinds of meaning-making would depend necessarily on the empirical details of individual features and particular speech communities, this idea deserves continued attention and vigorous investigative effort.

As far as I know, this is the first work to systematically explore subjective reactions to speakers using linguistic variables in different conditioning environments. As such, there are far many more questions raised here than answered. The first is to see if there are any other examples of meaningful variability whose semiotic value is affected by linguistic context. What makes a variant more or less likely to be affected by such context? There is also still much work to be done charting out the basic social meanings of creaky voice in different languages, cultures and discourse contexts.

There are a few elements of the experiment design in this chapter that have left data open for future development and analysis. There are two speakers in this experiment—due to the scale of the work I have treated speaker identity merely as a source of random variability, but with more stimuli and participants in a balanced design it would be fascinating to investigate differences in ratings between the two speakers, relative to how they are perceived by listeners in general. Persona-level expectations and inferences about speakers heavily color listeners’ evaluations of their socially meaningful language use (Podesva et al. [n.d.] Campbell-Kibler 2007, Pharao et al. 2014). The question of how deeply this effect interacts with the linguistic particulars—both the phonetic/linguistic characteristics of the feature itself

---

1 Although Gussenhoven’s account, which posits multiple such “codes” and sketches out a picture of how cross-linguistic variability in intonational meanings, could be derived from their interaction, it still adopts a basically ethological and essentialist view of the meaning of pitch, one which we should exercise extreme caution in addressing.
as well as the language structures that form the “envelope” of variation outlining it—is a logical next step for any sociolinguist interested in the sociophonetics of language and identity. Obviously there is room for bottom-up evidence from speech to set one’s expectations about a speaker (Staum Casasanto 2010). We might reasonably predict, however, that subtleties of discourse and linguistic context are more apt to inflect social meanings and pragmatic functions in situations where a priori, top-down, persona-level expectations are not highly determined. These might include encounters between strangers of similar social backgrounds, encounters where the evaluating participant is unfamiliar with the social landscape, or any other situation where systems of organized expectation (language ideologies, social scripts, etc.) are not strongly at play.

This is not just a proviso of “microsociolinguistic” interest. Though sociolinguistic variation studies has traditionally focused on variation in speech communities or other community-like collectives, there is increasing pressure from within and outside the field to acknowledge epochal shifts between the comparatively rigid social structures of archaic social formations (with very community-esque bundles of social relations and strong top-down pressure on persona-level identities and linguistic practice), through the partial reconfiguration of these relationships in so-called “high modernity,” and onward to their reputed breakdown in “late modernity,” where ideas of community and location have yielded a lot of ground to notions of network and movement. It is just such circumstances that compel us to reassess the importance of socially determined role identities and mediatized “characterological figures” (Agha 2007a,b).

Beyond these broader theoretical questions we have several other important questions we can ask, chief among them about the effects of the demographic and social traits of the raters themselves on their evaluative behavior. As the table in Appendix B lays out, raters came from a highly diverse array of national, linguistic, and social
backgrounds. Both mainland China and Taiwan have an official language standard based on Beijing pronunciation, but there may be significant differences in how linguistic variants are perceived by Taiwanese and mainlanders. Add to this diversity in age, dialect background, and even ethnic or subethnic identity (which is not information I collected here) and there is ample room for expansion of this model of research toward a fuller understanding of the social distribution of sociolinguistic perception.

I believe that to the extent linguistic context transforms the evaluations of individual forms, it will tend to do so not by merely switching one enregistered meaning ‘on’ and another one ‘off,’ but by affording a space for the relationship between competing schemata for computing meanings to change, allowing different (perhaps non-enregistered) interpretations to arise from concurrent, plausible evaluative frameworks.

In terms of the broader picture, these questions have implications for the description and organization of sociolinguistic and metalinguistic knowledge. The answers would help us sort out how deeply intertwined the social and pragmatic functions of a form are with individual speakers’ understandings of linguistic and discourse structure.
<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.83</td>
<td>4.63</td>
<td>2.79</td>
</tr>
<tr>
<td>position10</td>
<td>-0.21</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>questioncityvillage</td>
<td>0.66</td>
<td>1.22</td>
<td>0.93</td>
</tr>
<tr>
<td>questioneducated</td>
<td>0.14</td>
<td>0.70</td>
<td>0.41</td>
</tr>
<tr>
<td>questionenthusiastic</td>
<td>-0.51</td>
<td>0.06</td>
<td>-0.23</td>
</tr>
<tr>
<td>questioninterested</td>
<td>-0.24</td>
<td>0.33</td>
<td>0.06</td>
</tr>
<tr>
<td>questionknowledge</td>
<td>-0.03</td>
<td>0.53</td>
<td>0.24</td>
</tr>
<tr>
<td>questionrich</td>
<td>-0.07</td>
<td>0.49</td>
<td>0.21</td>
</tr>
<tr>
<td>questionsocialposn</td>
<td>-0.08</td>
<td>0.48</td>
<td>0.21</td>
</tr>
<tr>
<td>is.creakyTRUE</td>
<td>-0.23</td>
<td>0.36</td>
<td>0.07</td>
</tr>
<tr>
<td>position10:questioncityvillage</td>
<td>-0.61</td>
<td>0.08</td>
<td>-0.27</td>
</tr>
<tr>
<td>position10:questioneducated</td>
<td>-0.39</td>
<td>0.30</td>
<td>-0.05</td>
</tr>
<tr>
<td>position10:questionenthusiastic</td>
<td>0.17</td>
<td>0.86</td>
<td>0.52</td>
</tr>
<tr>
<td>position10:questioninterested</td>
<td>0.08</td>
<td>0.77</td>
<td>0.43</td>
</tr>
<tr>
<td>position10:questionknowledge</td>
<td>-0.38</td>
<td>0.31</td>
<td>-0.04</td>
</tr>
<tr>
<td>position10:questionrich</td>
<td>-0.33</td>
<td>0.36</td>
<td>0.02</td>
</tr>
<tr>
<td>position10:questionsocialposn</td>
<td>-0.47</td>
<td>0.23</td>
<td>-0.11</td>
</tr>
<tr>
<td>position10:is.creakyTRUE</td>
<td>-0.43</td>
<td>0.29</td>
<td>-0.08</td>
</tr>
<tr>
<td>questioncityvillage:is.creakyTRUE</td>
<td>-0.58</td>
<td>0.24</td>
<td>-0.16</td>
</tr>
<tr>
<td>questioneducated:is.creakyTRUE</td>
<td>-0.53</td>
<td>0.29</td>
<td>-0.11</td>
</tr>
<tr>
<td>questionenthusiastic:is.creakyTRUE</td>
<td>-0.12</td>
<td>0.69</td>
<td>0.28</td>
</tr>
<tr>
<td>questioninterested:is.creakyTRUE</td>
<td>-0.13</td>
<td>0.69</td>
<td>0.29</td>
</tr>
<tr>
<td>questionknowledge:is.creakyTRUE</td>
<td>-0.43</td>
<td>0.41</td>
<td>-0.01</td>
</tr>
<tr>
<td>questionrich:is.creakyTRUE</td>
<td>-0.45</td>
<td>0.37</td>
<td>-0.03</td>
</tr>
<tr>
<td>questionsocialposn:is.creakyTRUE</td>
<td>-0.56</td>
<td>0.26</td>
<td>-0.15</td>
</tr>
<tr>
<td>position10:questioncityvillage:is.creakyTRUE</td>
<td>-0.09</td>
<td>0.91</td>
<td>0.41</td>
</tr>
<tr>
<td>position10:questioneducated:is.creakyTRUE</td>
<td>-0.41</td>
<td>0.59</td>
<td>0.10</td>
</tr>
<tr>
<td>position10:questionenthusiastic:is.creakyTRUE</td>
<td>-1.04</td>
<td>-0.04</td>
<td>-0.54</td>
</tr>
<tr>
<td>position10:questioninterested:is.creakyTRUE</td>
<td>-0.96</td>
<td>0.04</td>
<td>-0.47</td>
</tr>
<tr>
<td>position10:questionknowledge:is.creakyTRUE</td>
<td>-0.48</td>
<td>0.53</td>
<td>0.02</td>
</tr>
<tr>
<td>position10:questionrich:is.creakyTRUE</td>
<td>-0.50</td>
<td>0.50</td>
<td>0.01</td>
</tr>
<tr>
<td>position10:questionsocialposn:is.creakyTRUE</td>
<td>-0.27</td>
<td>0.74</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table 5.13: Mixed effects model of all response values
Chapter 6

Conclusion

Exploring sociophonetic variability in creaky voice has brought us to several new understandings with regard to questions in sociolinguistics and the phonetics of Mandarin.

6.1 Linguistic Questions

This dissertation has confirmed that creaky voice quality in Mandarin is subject to constraints from both tone and the prosodic environment. That creaky voice would appear on tones whose primary cue is low pitch (tone 3) is consistent with other phonetic work on Mandarin [Belotel-Grenié & Grenié 1997], which has also found high rates of creaky voice for tone 3. I also replicated Belotel-Grenié and Belotel-Grenié’s finding that tone 4 also occasionally shows creaky voice; their numbers were much lower (18/56 syllables) for tone 4 than for tone 3 (40/51 syllables), and in my study, where tone 4 and tone 3 were both lower in spectral tilt (H1-A1 and H1-A2), the effect was lower in magnitude for tone 4.

The finding that neutral tone also often tends to be creaky is intriguing, and has not previously been reported, to my knowledge. The phonological status of neutral tone is a matter of considerable debate, with three main proposals available. One is that neutral tone syllables receive tone from the preceding syllable (a ‘spreading’ account, Yip 2002 181-182), and are possibly underlingly [-Upper] (a designation...
they share with tone 3 in Yip’s [1980] account). Another is an ‘interpolation’ account
(Shih 1987), where a neutral tone’s pitch features are interpolated phonetically from
the pitch of surrounding syllables. More recently, some have argued that neutral tone
has its own pitch target (van Santen et al. 1998), which enters into interpolation
or, alternatively, target approximation (the proposal of Chen & Xu 2006). Further
research could explore whether the phonatory characteristics, especially the timing of
phonatory events over the course of the syllable, provide more support to one model
of tone or another.

The interaction of tone and prosodic finality, especially in the apparent increased
breathiness of final neutral tone syllables deserves more attention as well. Not only
might it have something to offer ongoing debates about tone-intonation interactions
in Mandarin, but it could help complicate accounts of tone in Mandarin by integrating
the facts of phonation into the mix. It may also be of interest to scholars of other
languages where tone and phonation interact deeply, such as Wu Chinese (Cao &
Maddieson 1992), Burmese (Gruber 2011), and Green Mong (Andruski & Ratliff
2000).

More broadly, this dissertation has been a demonstration of the potential of socio-
phonetic work using mid-sized to large corpora of naturalistic production to contribute
to questions in phonetic and theoretical phonology, as well as to continue bringing
phonetic detail into studies of sociolinguistic variation.

6.2 The social meanings of creaky voice

One point which this study shows is that it may be useful to distinguish between the
positive content of a form’s meaning and the field in which that meaning is achieved.
While forms with highly enregistered metalinguistic or metapragmatic construals may
be described as ‘meaning something’ (having a specific indexical content) more than unenregistered forms do, that does not mean that we can say nothing about the semiotic consequences of using a form with fewer enregistered associations—like, perhaps, creaky voice. The effects which creaky voice had on evaluations of speakers were completely inconsistent across different linguistic contexts, but it only had an effect on two response dimensions—interest and enthusiasm. One explanation of this is that the indexical content of creaky voice is not fixed, but the field in which it operates is somehow limited to questions of determining speaker interest/enthusiasm.

6.2.1 Meaning as process

This leads us firmly into a view of social meaning as a kind of process, rather than a static product or structural element. This is well in keeping with contemporary work on indexicality and histories of enregisterment [Silverstein 2003, Agha 2007a]. A processual view helps in accounting for iconized and non-arbitrary form-meaning relationships in that it separates the culturally relative or contextually variable frameworks, ideologies, or schemata for generating meanings from the natural(ized) modality of the form-meaning relationship itself (based in resemblance, i.e. “iconic,” or otherwise).

Let’s use Silverstein’s (2003) analysis of /r/-lessness in New York’s Lower East Side as related to ideologies of orthographic “fullness” as an example. The ideology which generates meaning here is one that compares phonetic forms to orthographic forms. The modality of the relationship specified is iconic, but since we have made explicit that the language ideology is culturally relative, we do not have to be surprised when this particular form-meaning relationship does not recur in other speech communities.

1A question generated by this approach is how this ideological construal of /r/-lessness, which relies on a general “folk theory” of the sign and also leads to stigma and “linguistic insecurity,” does not overapply and end up stigmatizing other cases where the orthographic
The basic issue in admitting iconization into the fold of meaning-making processes is, as De Cuypere (2008) has framed it, “limiting” it. One way of doing so—hardly a radical one (Irvine 2005)—is to recall that many linguistic forms that are meaningful in iconic modalities are iconic in more than one way. Furthermore, the “grounds” on which they are iconic—that is, the sign-objects which they may be taken to resemble—can themselves be constituted by a variety of means. One is ideology, as in the already well-discussed case of Silverstein’s interpretation of /r/-lessness. Another example of this is the iconization of clicks in Nguni to reflect the (ideologically invented and sustained) otherness of the Khoi languages from which they were borrowed (Irvine & Gal 2000). Discourse dynamics in interactional time can also provide a basis for iconic meaning construals, as in the iconic relation between the order of clauses in a narrative and the temporal order of events they refer to (Jakobson 1963; Haiman 1980). Finally, as the findings of this dissertation support, language structure itself can provide iconic grounds for constructing relationships of similarity.

6.2.2 How might language intervene in social meaning processes?

My proposal in chapter 5 is that creaky voice can potentially be interpreted according to two schemata in Beijing Mandarin. Prosodic finality invites the listener to expect low pitch targets, in which context creak can be heard as a super-low implementation of a low pitch target. Exceeding predicted performance in this way can be semiotized as “interest” or “enthusiasm,” perhaps in the same way we might expect increased loudness to perform a similar function. Medially, where these expectations of low signal exceed what the phonetic signal provides. In other words, why don’t 60’s New Yorkers stigmatize “silent e” in words like torque? One solution is to throw out Silverstein’s theory as half-baked; another might involve stipulating that the ideology is invoked only in cases of /r/-lessness, though the specifics of this proposal would need to be worked out at length.
pitch targets are not at play, creak may be heard as “lax” or “low energy” and thus be used to read the speaker as relatively lacking in interest and enthusiasm.

Language structure enters into this picture as a set of hearer expectations. The variable production of some feature is overlaid against these expectations, and, in some cases, evaluated according to the schemata that these expectations make relevant. More instances of variability in meaning according to linguistic context (and also of lack thereof!) should be sought out. In comparative perspective it may become clearer what role language and/or language structure play in influencing the meanings of variability in speech and language.

How might the findings of this dissertation play into further investigations of voice quality? First it should be determined whether interest and enthusiasm are indeed domains on which creakiness can generally signify. This could be done by using more elaborate response instruments in an overall similar design, looking at mass-mediated depictions of “enthusiasm” (or lack thereof), or by examining interactionally valid displays of “interest” or “enthusiasm.”

Moreover, the links between possible sociophonetic differences in the creak of “urban male Beijingers” and “rural female Beijingers” should be investigated. Do mass-mediated depictions of urban and rural subjects treat creak differently? Depict it in different stylistic contexts? Show gender differentiation?

6.3 Methodological issues

This dissertation’s methodologies are largely in the vein of well-established studies of sociolinguistic variation. The use of the sociolinguistic interview as a source for data on the social and linguistic distribution of the variable in question originates with the Lower East Side study (Labov 2006 [1966]) (with antecedents in Labov 1972 [1963]),
which also pioneered investigations into listeners’ subjective reactions to variation. The specific matched guise methodology that I pursued also has roots in the 1960’s (Lambert et al. 1966) and has been refined and adapted by many sociolinguists (Giles & Bourhis 1976; Rickford 1985; Campbell-Kibler 2007; Podesva et al. n.d.).

Nevertheless, the innovations of method this dissertation has introduced are worth noting, especially for the indications they make for future work.

6.3.1 “Big-ish data” and economies of method

Labov, Rosenfelder & Fruehwald (2013) was a big moment in sociolinguistics for a few reasons. Besides being one of the most detailed statements on the empirical study of language change in progress to date and appearing in the flagship journal of linguistics, it was also one of the first major pieces of research on linguistic variation to use automatic forced alignment (Yuan & Liberman 2008) to extract audio for measurement.

The major advantages of using forced alignment to look at large (or even just large-ish) amounts of data are that it is cheap, fast, and easy. When you use forced alignment, after data collection itself and orthographic transcription are accounted for, the other labor-intensive stages of analysis do not increase in difficulty with the size of the data. The trade-off in data quality, meanwhile, is comparatively small, and with enough data, outlier measurements generated from misaligned segments can be handled by using robust measures of central tendency or probabilistic methods for identifying and excluding them from the data. As the use of forced alignment increases, as well as in future iterations of this work, it will be important to incorporate such practices into the default workflow associated with forced alignment.

I was there first, of course—the last listed author on an ADS talk (Podesva et al. 2007) that used forced alignment to study the speech of former Secretary of State Condoleezza Rice.
The relationship of forced alignment to the so-called “third wave” of variation studies—attention to how meaning arises in context—has yet to be fully hammered out, beyond helping establish wider background patterns against which the drama of meaning in interaction unfolds. Podesva, Brenier, Hall-Lew, Lewis, Callier & Starr (2007) found forced alignment useful for creating a so-called “sociophonetic profile” of a speaker, which included information on speech in different settings. But for other analytic contexts, such as the use and distribution of variants in particular discourse contexts with inherently low Ns, the usefulness of forced alignment at the stage of measurement and analysis is low, since precision as well as attention to sequential details are called for. Even in contexts like this, however, measurements obtained through forced alignment can be useful for identifying zones of interaction where temporal patterns of variation attract attention (sometimes called “style clusters”), or where phonetically “extreme” variants (which are particularly illustrative in investigations of meaning; see Podesva 2011) may be located. Even after wading through outliers, finding genuinely extreme tokens of a gradient measurement dimension will take less time when the way has been cleared by forced alignment.

6.3.2 Experimental work in context

The experimental component of this dissertation responds to multiple areas of concern. One is simply to uncover the indexical meaning potentials of creaky voice, for which there is little extant work in Mandarin (and only slowly emerging work in English). From a linguistic standpoint, modifying prosodic and tonal context to introduce creak in different linguistic environments helps us understand whether linguistic context changes a social index’s meaningfulness to listeners.

The result—that at least prosodic context matters in determining the meaning of creak—is a new finding; however, it will be important in subsequent work to see if
other methodologies triangulate to similar conclusions. Comparisons between different naturalistic interactional situations could be made, tallying up levels of creaky voice quality in different linguistic environments, a la (Podesva 2007). Other voice qualities could be investigated using similar paradigms to see if the same patterns resurface. Different methods of introducing voice quality—for instance, using artificial synthesis rather than splicing—could be used to make analogous stimuli. The variation in stimulus content between prosodic and tonal contents might also be reduced by using more similar sentences. Conversely, more in the vein of Campbell-Kibler (2007), we might go in the direction of having more speakers, with slightly more “distinctive” speech passages, so as not to cancel out the meanings that arise when speakers’ identities are more “marked” and their evident affective investment in the speech is higher. Of course, above all, replicating the study with a higher number of participants is also key to enhancing our confidence in the tentative findings reported here.

More importantly, though, findings regarding the “meaning” of creaky voice quality will ideally be enhanced by investigations grounded in ethnographic study of particular sites, speech communities, and/or communities of practice. The particular linguistic resources available to speakers have often been shown to take on particular meanings (often as higher-order indexical construals) in particular community contexts (Rickford 1986; Eckert & McConnell-Ginet 1992; Moore & Podesva 2009; Eckert 2012). One suggestion which arises perennially is to conduct experiments akin to the one done in this dissertation in ethnographic contexts where the researcher is involved. The structure of sociolinguistic fieldwork as ordinarily practiced, where distributional analysis typically takes place “away” from the field, often when it is difficult for the researcher to return, makes it hard to do this, but certainly not impossible. Well-studied variables, the general or predicted shape of whose (socio)linguistic distribution can be tentatively sketched out in advance, might allow for faster iter-
ation of experimental designs, perhaps fast enough to be implemented in the field. Deep and lasting personal ties between fieldworkers and informants, or at least ties that can be in some sense “renewed” months or years onward, would also facilitate such inquiries.

In general, though the experimental approach adopted here is, with proper controls, sometimes the only way to properly “diagnose” precisely what a variant is adding to the semiotic mix of an utterance, a processual, context-sensitive view of meaning strongly indicates both deep ethnographic involvement and close attention to the sequential facts of discourse (the “co-text”). Balancing the requirements and affordances of all of these approaches is a tricky task, but will bring us much closer to a realistic understanding of sociophonetic variation in context.

6.4 What we still want to know

Voice quality is an appealing topic for both its phonetic and social richness. Even within just creaky voice, the four different spectral tilt measures we used, along with my hand-coding, each gave a different picture of its social and linguistic distribution. Distilling out, to the extent possible, what of this variance is linguistically or socially meaningful is a difficult task, but one it is worthwhile to pursue.

Voice quality seems to play a key role in achieving distinction among characterological figures in mass-mediated discourse (Mendoza-Denton 2011; Callier 2012), as well as serving key functions in narrative and ongoing discourse (Podesva 2013; Sicoli 2010). In particular, I feel that work remains to be done in interpreting and breaking apart the dimensions of voice that are used to signify otherness in mediated characterological representations—for example, ethnic others in video games in the Grand Theft Auto series, as in Mendoza-Denton’s work, or threats to the estab-
lished order as in my own work on Chinese television. How, if at all, are these linked to non-characterological uses of non-modal voice quality in ordinary discourse? And can sociolinguists contribute meaningfully to broader dialogues about the mediated representation of “others,” particularly marginalized others?

And if, as appears to be happening in the United States, voice quality itself comes to be an object of mass-mediated explicit metalinguistic evaluation, as discussed in the introduction, how can we, as linguists, a) promote scientific understanding of the underlying physical and anatomical realities—for instance, to depathologize “vocal fry”—and b) actively oppose the weaponization of “vocal fry” (as happened in the case of “uptalk”) in attempts to make a spectacle (and morality play) out of (young, white) women’s “speech patterns”? One could ask similar questions in the Chinese context, not just about the representation of women, but also about multimodal mediated depictions of 外地人 waidiren ‘outlanders, people from elsewhere [in China]’ and 农民工/民工 nongmingong/mingong ‘migrant workers’ in the context of China’s precipitous urbanization. Entering into these discourses would be fighting an uphill battle, and where voice quality is specifically concerned, the literature requires perhaps a bit more maturation. But given voice quality’s importance in everyday discourse, another “vocal fry” flap could happen at any moment. Public engagement should not be far from the mind of any sociolinguist, and the twin tasks of data collection and theory-building which this dissertation takes on are necessary groundwork for this kind of usefulness.
Appendix A

Pinyin-to-ARPABET mapping for P2FA

<table>
<thead>
<tr>
<th>Pinyin</th>
<th>IPA</th>
<th>Pseudo-ARPABET</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>tc</td>
<td>JH</td>
</tr>
<tr>
<td>Q</td>
<td>tcʰ</td>
<td>CH</td>
</tr>
<tr>
<td>X</td>
<td>c</td>
<td>SH</td>
</tr>
<tr>
<td>C</td>
<td>tsʰ</td>
<td>T S</td>
</tr>
<tr>
<td>Z</td>
<td>ts</td>
<td>D Z</td>
</tr>
<tr>
<td>ZH</td>
<td>tʂ</td>
<td>JH</td>
</tr>
<tr>
<td>H</td>
<td>h~x</td>
<td>HH</td>
</tr>
<tr>
<td>all others</td>
<td></td>
<td>identical to pinyin</td>
</tr>
</tbody>
</table>

Table A.1: Mapping from pinyin to pseudo-ARPABET for P2FA, pinyin initials only.
<table>
<thead>
<tr>
<th>Pinyin</th>
<th>IPA</th>
<th>Pseudo-ARPABET</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>AA1</td>
</tr>
<tr>
<td>AI</td>
<td>ai</td>
<td>AA1 IY1</td>
</tr>
<tr>
<td>AO</td>
<td>au</td>
<td>AA1 UW1</td>
</tr>
<tr>
<td>AN</td>
<td>an</td>
<td>AA1 N</td>
</tr>
<tr>
<td>ANG</td>
<td>ang</td>
<td>AA1 NG</td>
</tr>
<tr>
<td>O</td>
<td>o</td>
<td>AO1</td>
</tr>
<tr>
<td>ONG</td>
<td>uŋ</td>
<td>OW1 NG</td>
</tr>
<tr>
<td>OU</td>
<td>ou</td>
<td>OW1</td>
</tr>
<tr>
<td>E</td>
<td>e</td>
<td>AH1</td>
</tr>
<tr>
<td>EI</td>
<td>ei</td>
<td>EY1</td>
</tr>
<tr>
<td>EN</td>
<td>en</td>
<td>AH1 N</td>
</tr>
<tr>
<td>ENG</td>
<td>en</td>
<td>AH1 NG</td>
</tr>
<tr>
<td>ER</td>
<td>er</td>
<td>AA1 R</td>
</tr>
<tr>
<td>I</td>
<td>i/i/ ui</td>
<td>Y IY1</td>
</tr>
<tr>
<td>IA</td>
<td>ja</td>
<td>Y AA1</td>
</tr>
<tr>
<td>IAN</td>
<td>jen</td>
<td>Y EH1 N</td>
</tr>
<tr>
<td>IAO</td>
<td>jou</td>
<td>Y OW1</td>
</tr>
<tr>
<td>IE</td>
<td>je</td>
<td>Y EH1</td>
</tr>
<tr>
<td>IU</td>
<td>jou</td>
<td>Y OW1</td>
</tr>
<tr>
<td>IANG</td>
<td>jʊŋ</td>
<td>Y AA1 NG</td>
</tr>
<tr>
<td>IN</td>
<td>in</td>
<td>Y IY1 N</td>
</tr>
<tr>
<td>ING</td>
<td>in</td>
<td>Y IY1 NG</td>
</tr>
<tr>
<td>IONG</td>
<td>qʊŋ</td>
<td>Y UH1 NG</td>
</tr>
<tr>
<td>U</td>
<td>u</td>
<td>UW1</td>
</tr>
<tr>
<td>UA</td>
<td>wa</td>
<td>W AA1</td>
</tr>
<tr>
<td>UO</td>
<td>wɔ~ wɔ́n</td>
<td>W AO1</td>
</tr>
<tr>
<td>UI</td>
<td>wei</td>
<td>W EY1</td>
</tr>
<tr>
<td>UAI</td>
<td>wai</td>
<td>W AA1 IY1</td>
</tr>
<tr>
<td>UAN</td>
<td>wan</td>
<td>W AA1 N</td>
</tr>
<tr>
<td>UN</td>
<td>wɛn</td>
<td>W AH1 N</td>
</tr>
<tr>
<td>UANG</td>
<td>wæŋ</td>
<td>W AA1 NG</td>
</tr>
<tr>
<td>U/U (post-palatal)</td>
<td>ʊ</td>
<td>IY1</td>
</tr>
<tr>
<td>ÜE/UE (post-palatal)</td>
<td>ʊəœ</td>
<td>Y EH1</td>
</tr>
<tr>
<td>UAN (post-palatal)</td>
<td>ʊən</td>
<td>Y EH1 N</td>
</tr>
<tr>
<td>UN (post-palatal)</td>
<td>ʊn</td>
<td>Y AH1 N</td>
</tr>
</tbody>
</table>

Table A.2: Mapping from pinyin to pseudo-ARPABET for P2FA, pinyin finals only.
## Appendix B

### Matched guise experiment participants and materials

<table>
<thead>
<tr>
<th>userid</th>
<th>mainlg</th>
<th>birthplace</th>
<th>gender</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>26001</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Female</td>
<td>26-35</td>
</tr>
<tr>
<td>27001</td>
<td>Putonghua</td>
<td>Hubei</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>28003</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>29002</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>29003</td>
<td>Putonghua</td>
<td>Jiangsu</td>
<td>Female</td>
<td>26-35</td>
</tr>
<tr>
<td>30002</td>
<td>Putonghua</td>
<td>Shanghai Municipality</td>
<td>Female</td>
<td>46-55</td>
</tr>
<tr>
<td>30003</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Male</td>
<td>26-35</td>
</tr>
<tr>
<td>32001</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Male</td>
<td>26-35</td>
</tr>
<tr>
<td>35001</td>
<td>Putonghua</td>
<td>Hubei</td>
<td>Female</td>
<td>26-35</td>
</tr>
<tr>
<td>37001</td>
<td>Mandarin</td>
<td>Henan</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>38001</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Male</td>
<td>26-35</td>
</tr>
<tr>
<td>39001</td>
<td>Putonghua</td>
<td>Fujian</td>
<td>Female</td>
<td>26-35</td>
</tr>
<tr>
<td>41001</td>
<td>Putonghua</td>
<td>Henan</td>
<td>Male</td>
<td>18-25</td>
</tr>
<tr>
<td>43001</td>
<td>Putonghua</td>
<td>Hubei</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>44001</td>
<td>Putonghua</td>
<td>Cantonese</td>
<td>Male</td>
<td>18-25</td>
</tr>
<tr>
<td>46001</td>
<td>Putonghua</td>
<td>Shanghai</td>
<td>Female</td>
<td>18-25</td>
</tr>
<tr>
<td>46002</td>
<td>Putonghua</td>
<td>Beijing</td>
<td>Male</td>
<td>26-35</td>
</tr>
<tr>
<td>49001</td>
<td>Wu Dialect</td>
<td>Zhejiang</td>
<td>Female</td>
<td>26-35</td>
</tr>
<tr>
<td>52001</td>
<td>Putonghua</td>
<td>Cantonese</td>
<td>Female</td>
<td>26-35</td>
</tr>
</tbody>
</table>

Table B.1: Participant information
<table>
<thead>
<tr>
<th>character</th>
<th>pinyin</th>
<th>tone</th>
<th>sentence</th>
<th>position</th>
<th>spkr1</th>
<th>spkr2</th>
</tr>
</thead>
<tbody>
<tr>
<td>万</td>
<td>wan</td>
<td>4</td>
<td>这几天他赚了十一万</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>凳</td>
<td>deng</td>
<td>4</td>
<td>小木板凳放在房子里了</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>凳</td>
<td>deng</td>
<td>4</td>
<td>我选择了左边的小木凳</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>到</td>
<td>dao</td>
<td>4</td>
<td>我问小华究竟有没有到</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>困</td>
<td>kun</td>
<td>4</td>
<td>夏天的午后人容易犯困</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>大</td>
<td>da</td>
<td>4</td>
<td>我家种的西瓜又圆又大</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>奉</td>
<td>feng</td>
<td>4</td>
<td>舍利子被各大寺庙供奉</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>展</td>
<td>zhan</td>
<td>3</td>
<td>国际会展中心在哪里呢</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>展</td>
<td>zhan</td>
<td>3</td>
<td>她决定在首都举办画展</td>
<td>10</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>岛</td>
<td>dao</td>
<td>3</td>
<td>我们去岛上参加摄影赛</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>岛</td>
<td>dao</td>
<td>3</td>
<td>珠海市附近有很多小岛</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>慢</td>
<td>man</td>
<td>4</td>
<td>这破车怎么跑得这么慢</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>慢</td>
<td>man</td>
<td>4</td>
<td>进步太慢只能接受批评</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>战</td>
<td>zhan</td>
<td>4</td>
<td>为自己的幸福生活而战</td>
<td>10</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>打</td>
<td>da</td>
<td>3</td>
<td>小时候王建斌经常挨打</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>打</td>
<td>da</td>
<td>3</td>
<td>张小明打了一辆出租车</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>我</td>
<td>zhao</td>
<td>3</td>
<td>钥匙落在树林里不好找</td>
<td>10</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>捆</td>
<td>kun</td>
<td>3</td>
<td>刚买的火柴是十捆一捆</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>捆</td>
<td>kun</td>
<td>3</td>
<td>我的一捆柴被人背走了</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>捧</td>
<td>peng</td>
<td>3</td>
<td>刘德华在中国非常受捧</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>捧</td>
<td>peng</td>
<td>3</td>
<td>把杯子捧在手中有点烫</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>斩</td>
<td>zhan</td>
<td>3</td>
<td>这次的结果是囚犯被斩</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>晚</td>
<td>wan</td>
<td>3</td>
<td>他们俩都觉得相见恨晚</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>满</td>
<td>man</td>
<td>3</td>
<td>一屋子满是没洗的衣服</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>满</td>
<td>man</td>
<td>3</td>
<td>这事让大家感到很不满</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>照</td>
<td>zhao</td>
<td>4</td>
<td>中国非常流行拍婚纱照</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>站</td>
<td>zhan</td>
<td>4</td>
<td>下午我按时到达了车站</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>站</td>
<td>zhan</td>
<td>4</td>
<td>公交车站旁有事故发生</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>等</td>
<td>deng</td>
<td>3</td>
<td>他没有等到小明就走了</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>等</td>
<td>deng</td>
<td>3</td>
<td>参加会议的有李玉钱等</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>讽</td>
<td>feng</td>
<td>3</td>
<td>冷嘲热讽会被他人瞧不起</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>讽</td>
<td>feng</td>
<td>3</td>
<td>小明受尽了旁人的嘲讽</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>道</td>
<td>dao</td>
<td>4</td>
<td>我春游走的是羊肠小道</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table B.2: Stimulus information
BIBLIOGRAPHY


BABEL, MOLLY; McGUIRE, GRANT; KING, JOSEPH & SATTERWHITE, ALYSSA (2011). Vocal aesthetics, prototypicality, and stereotypicality.


CAMPBELL-KIBLER, KATHRYN (2007). Accent, (ING), and the social logic of listener perceptions. American Speech, 82(1).


Labov, William; Rosenfelder, Ingrid & Fruehwald, Josef (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. Language, 89(1): 30–65. ISSN 1535-0665. doi:10.1353/


URL http://americanspeech.dukejournals.org/content/86/1/69.abstract


PHARAO, NICOLAI; MAEGARD, MARIE; MØLLER, JANUS SPINDLER & KRISTIANSSEN, TORE (2014). Indexical meanings of [s+] among Copenhagen youth:
Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, 43(1).


van Santen, Jan PH; Möbius, Bernd; Venditti, Jennifer & Shih, Chilin (1998). Description of the Bell Labs intonation system. In *Proceedings of Third*


Xu, Xiaoying; Ye, Weiping; He, Yi; Zhao, Xinru & Callier, Patrick R. (2012). Phonetic and social constraints on creaky voice—a research project based on the BNUPSC corpus. Tianjin, China: Presented at 8th International Conference on Chinese Sociolinguistics.


—— (2007b). "Jiang Shopping Jinxing Daodi!": Linguistic Innovation and Social Distinction in Chinese Television Medium. In *Hallett, Taryne; Floyd, Simeon; ...
Oshima, Sae & Shield, Aaron (eds.), Texas Linguistic Forum, vol. 50. Austin, TX.
