The Role of Socioindexical Expectation in Speech Perception

by

Kevin B. McGowan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in The University of Michigan
2011

Doctoral Committee:

Professor Stephen P. Abney, Co-chair
Professor Patrice Speeter Beddor, Co-chair
Professor Julie Boland
Associate Professor Robin Queen
Associate Professor Benjamin Munson, University of Minnesota

I dedicate this thesis to my sister and my mother who shaped me, to my wife and my daughter who supported me and to the thousands of peanut M&Ms who selflessly gave their lives that I might write.

ACKNOWLEDGEMENTS

It is impossible to heap sufficient praise and thanks on Pam Beddor. She is a supernaturally generous adviser, teacher, boss, colleague and friend. Pam taught me my first linguistics class as an undergraduate, modeled what it means to be a careful, dedicated experimentalist and has read all of my drafts of everything I've done and filled every margin with tiny, invaluable comments. She wrote my reference letters for graduate school and fellowships and countless job applications, took me on a pilgrimage to Haskins and has proven to be the most astonishingly giving and supportive adviser a fellow could ask for.

Steve Abney is a patient and insightful mentor and his careful attention to my research has improved it tremendously. Steve warned me on my first day of graduate school that this would be difficult. I remember nodding gravely and telling him I knew. I did not know.

Thanks to my committee members: Julie Boland, Robin Queen and Ben Munson. I will never understand what possessed me to assemble five of the most erudite and intimidating people I could find, but they did it gently, helpfully and openhandedly. The various errors, omissions and blunders I've playfully hidden –easter egg like– around the thesis are entirely my own doing.

I am deeply grateful to Carson Maynard of the University of Michigan's English Language Institute for his help understanding what difficulties native Mandarin and Korean speakers have learning English. I owe similar thanks to the native Mandarin speakers in Michigan Linguistics who have allowed me, often with their knowledge and consent, to analyze their speech.

I owe many, many thanks to Keith Johnson and everyone at UC Berkeley for letting me take over their lab for a week, helping me recruit subjects and for letting me play with all of the cool toys in the phonology lab.

Ed & Sue Kusnir have donated hundreds of afternoons watching Nora for us and have supported us in more ways than I can count. No one could wish for better parents-in-law or grandparents.

I would never even have gone to college, much less earned a PhD, if my big sister hadn't literally dragged me there. Thank you, Colleen.

My entire family, especially Brian, Jim, Brien, Brien, Erin, Beth, and my Mom, have worked diligently to keep my natural arrogance in check these last few decades. You lot are the only people who will have to call me 'Dr.' now; it would be best if you could avert your eyes and bow slightly when saying it. You're welcome.

Sam Epstein, you're not getting any more posters outta me, buddy. No way. Actually, come to think of it, you can call me 'Dr.' too.

Sylvia, Sandie and Kristen in the Linguistics office have saved my bacon more often than I care to commit to print (even if it *is* print no one will ever read).

I have had some wonderful teachers and mentors at Michigan: Sally Thomason, Andries Coetzee, John Lawler, Dragomir Radev, Richard W. Bailey, Lesley Milroy and Rosina Lippi-Green. Each one of these people made a huge impression on me –usually at just exactly the moment I needed them to.

It was an honor to share a cohort with Lauren Squires, Miyeon Ahn, and Miki Obata. Thanks for setting such a brisk pace, Lauren, and for teaching me the word 'indexical'.

Thank you to my fellow students Cati Fortin, Chris Odato, Rob Felty, Erika Alpert, Erica Beck, Terry Sczymanski, Damon Tutunjian and Joseph Tyler. I am happy to call you my friends.

I will especially miss (I already miss) my lab mates Anthony Brasher, Susan Lin, Jonathon Yip, Cameron Rule, Harim Kwon and honorary lab mates David Medeiros and Anna Babel. I absolutely could not have finished this dissertation without Anthony. Susan made every day in the lab a nerdy, hard-working joy and finished all of my sentences I was too tired or distracted to finish. If it weren't for Cameron I'd be whimpering on the side of a road somewhere in Kentucky trying to bribe someone to drive my rental truck to Texas. David, you're a freaking genius and the tongue twisters project is a great idea (no matter what the audience at LSA thought). I think Anna has put nearly as much thought and energy into this dissertation as I have. You are a great friend and colleague, Anna. Thanks for all of your ideas, discussion and encouragement.

I owe great thanks to Sandy, Wes, Jojo, Liam, Tony, Kasthuri, Gabi, Sean, Andrew, Andy, Willie and everyone who made my previous life so difficult to give up. And many thanks to my North family –Markus, Hans, Joe, Christine and Al– for teaching me a new kind of confidence.

Thanks to Robert Englebretson for his supportive correspondence during my summer of writing and to Nancy Niedzielski for trusting that I could get this thing done.

Finally, I have to thank my little family. Jen, you have carried us on your back, dragged me to my laptop, kept us clothed and fed, and produced an amazing little person while we've been in graduate school. Thank you. I love you. I think I've had my share of the cookie for a while. Nora Evelyn McGowan, you are a delight. You are the song in my heart. You are the Phial of Galadriel. You are sunshine in pink sneakers. You are the Typhoid Mary of happiness. Seeing the world reflected in your glittering, blue eyes has made everything brighter, more fascinating, deeper, more worth doing and a whole lot more fun. With every fiber of my being and every ounce of my energy I adore you. However, you were absolutely no help at all with this blasted dissertation.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

LIST OF APPENDICES

Appendix

ABSTRACT

The Role of Socioindexical Expectation in Speech Perception

by

Kevin B. McGowan

Chairs: Steven P. Abney & Patrice S. Beddor

Listeners extract cues to speaker identity from the speech stream. Recent evidence suggests that listeners will also perceive these socioindexical cues, even if absent, when primed to expect them. Most researchers interpret these findings as evidence for exemplar models of speech perception (Niedzielski, 1999; Hay et al., 2006b; Staum Casasanto, 2009a). At least one early line of research, however, attributes the influence of socioindexical knowledge on speech perception to listeners' negative bias (Rubin, 1992).

A series of three experiments with experienced and inexperienced listeners investigates the use of socioindexical expectation during speech perception. The first experiment, a yes/no accent identification task, reveals that listeners, whether experienced or inexperienced with Chinese-accented English, are capable of judging the authenticity of a non-native accent. Experienced listeners are significantly more accurate and inexperienced listeners are significantly more likely to rate an imitated accent as authentic –suggesting they depend more heavily on stereotypical features.

Experiment 2 addresses whether listeners can use socioindexical expectations to *enhance* speech perception. Both experienced and inexperienced listeners were significantly better at transcribing Chinese-accented sentences in noise when presented with an Asian face than when presented either with a silhouette or a Caucasian face. This result suggests that the negative bias hypothesis can not be correct; listeners can use socioindexical cues to enhance speech perception.

Experiment 3 used eye-tracking to investigate the time course of the influence of socioindexical expectation. Inexperienced listeners hearing a Standard American English voice showed no significant difference in fixation latencies when presented either with an Asian or Caucasian face. Listeners shown an Asian face showed significantly longer dwell times to the target image late in the trial, however. This result reinforces the finding that the negative bias hypothesis can not be correct but is also not consistent with an exemplar based account in which socioindexical expectation pre-activates groups of socially labeled exemplars (Johnson, 2006).

These results point to the need for more natural –more linguistic– tasks in the investigation of socioindexical speech perception and the need to look closely at time course to better understand the role of socioindexical expectation in speech perception.

CHAPTER I

Introduction

1.1 Introduction

This dissertation is a study of listeners' ability to use socioindexical information during speech perception. The term 'speech perception' here is construed to include the receipt of sensory information by the peripheral sense organs, the analysis of sensory input into possible sub-lexical abstract representations, the selection of potential lexical candidates given the transduced sensory input and the ultimate retrieval and recognition of a lexical item or items.

This work should be seen as part of a larger, more general project to understand the role of listener expectations during speech perception and word recognition. An assumption of this framework is that social knowledge and stereotypes influence the processing of fine phonetic detail and the retrieval of categorical lexical representations from memory. It is anticipated that this influence will not differ substantially from other kinds of listener expectation, including: expectation due to coarticulation, expectation of allophonic variation due to segmental or syllabic context, expectation created by input to other sensory modalities, expectation due to semantic or pragmatic knowledge of the speech situation, and even anticipation due to the phonemic categories and lexicon of one's native language(s).

In this opening chapter I will attempt to situate recent work on what has been called sociophonetics within the larger framework of speech perception research in general. The goal here is to show that disparate findings over decades of speech perception research seem to converge on the observation that the mapping of acoustic stimulus to categorical mental representations is far more malleable –much more susceptible to influence from segmental context, multi-modal sensory input, semantic knowledge, etc.– than one might initially imagine (cf. Liberman, 1996).

Socioindexical information, whether visual, auditory or salient in the communicative context, includes cues to such features as speaker gender, age, socio-economic status, sexual orientation, ethnicity, regional background, and markers of individual identity (Foulkes, 2010). These 'non-linguistic' aspects of the speech signal have long been acknowledged by speech perception researchers (e.g. Peterson and Barney, 1952; Ladefoged and Broadbent, 1957) but have only recently become an active object of research –particularly in the laboratory phonology community (Croot, 2010).

1.1.1 The Exemplar Hypothesis

This surge of interest in the perception of socioindexical aspects of speech is widely, if not universally, attributed to the rise of exemplar models in cognitive psychology and related disciplines (e.g. Hintzman, 1986; Nosofsky, 1986; Goldinger, 1998; Nosofsky and Zaki, 2002; Labov, 2006; Hay et al., 2006b; Foulkes, 2010; Munson, 2010). I will argue that this attribution is both too specific and possibly premature given the available evidence. "Too specific" because, with a few exceptions (e.g. Johnson, 1997, 2006; Pierrehumbert, 2001), linguists' interest in exemplar models is largely limited to those models' accommodation of the learning and use of fine phonetic detail and indexical knowledge. The neural substrate of this system and even the particulars of the learning algorithm are generally secondary to this affordance of learning (though see Goldinger (2007) for an exploration of the neural substrate of exemplar models).

I will also argue that interpreting most existing work in socioindexical perception as evidence supporting exemplar models (specifically as evidence of stored episodic traces labeled with both linguistic and social category information) is somewhat premature. Much of this interpretation rests on the critical assumption that listeners' early judgments of socioindexical phonetic detail will remain constant throughout a listening task and that these judgments will not be shifted, supplanted or, indeed, created by later, higher level cognitive processes. The tasks that have been used do not, in general, explore the time course of the influence of socioindexical knowledge on perception, but the analysis of these results as evidence for exemplar models *does* make a time course claim. This claim, specifically, is that socioindexical features are part of the episodic traces in the listeners' mental lexicon and that socioindexical influence on lexical access will happen quite early in the processing of speech. A recurring theme in this dissertation will be the need for more natural linguistic tasks with more direct on-line assessments of listeners' behavioral responses to stimuli to better evaluate hypotheses like this one.

Many studies have clearly shown that listeners' performance on listening tasks can be altered by altering their beliefs about the identity of the speaker. In one classic study, Niedzielski (1999) played recordings of a native Detroit, Michigan speaker for two groups of Detroit listeners. There is a widely held belief among Michigan speakers that their variety of English is 'unaccented' or identical to Standard American English (Niedzielski, 1995, 1997). Though both groups heard identical recordings, one group was led to believe the speaker was Canadian while the second group was led to believe the speaker was a fellow Detroiter. When asked to match what they'd heard to tokens with resynthesized vowels, the listeners who believed the speaker to be a fellow Detroiter were significantly less likely to choose an accurately raised and tensed [æ] or an accurately backed and lowered [ɛ] consistent with Detroit speakers' general participation in the Northern Cities Vowel Shift (Milroy and Gordon, 2003). Instead, listeners preferred resynthesized vowels unlike what they'd heard but closer to those typical of Standard American English.

When Niedzielski and those who have since reused the 1999 task (e.g. Hay and Drager, 2010) interpret listener responses in terms of the use of phonetic detail, they clearly assume that classification judgments made some moments after hearing a target word will faithfully recapitulate any indexical perceptions made while processing the speech signal and while the phonetic details of the target signal were active in the listeners' attention.

1.1.2 The Negative Bias Hypothesis

Additionally, I will consider the hypothesis that the results of socioindexical speech perception experiments may be attributable merely to negative bias on the part of the listener.

Rubin (1992) played identical mini-lectures to undergraduate students who were shown a photograph of either an Asian or Caucasian graduate student instructor whom they were led to believe was the speaker on the audio tape. Rubin found that listeners who saw the Asian photograph perceived the voice to be more strongly accented. These listeners also tended to have lower scores on a comprehension test –although this different was not statistically significant.

Listeners who believed the instructor to be Asian tended to retain particular lexical items less well. Rubin interprets this finding as evidence of negative social bias on behalf of the listeners –specifically, due to a lack of homophyly. These negative attitudes toward Asian instructors lead, Rubin argues, to a communicative breakdown in the classroom separate from any legitimate claims about difficulty understanding a non-native speaker's accented English. Regardless of how hard the non-native instructor may have worked to achieve a native-like English accent (indeed, even if this effort is completely successful), listener bias will still result in perceptions of accentedness and reduced perceptibility.

This interpretation of Rubin (1992) is later endorsed and extended by Lippi-Green (1997). Lippi-Green describes these results in the most stirring terms. She introduces the concept of ‘communicative burden’ which is conceptually similar, though at a social level, to the H&H theory of speech perception (Lindblom, 1990). Listeners and speakers are engaged in the shared act of communicating. Speakers control, to the best of their ability, how much energy they expend producing speech that is maximally clear to the listener, but listeners are not merely passive receivers. Listeners control how much energy *they* are willing to expend in decoding the signal, resolving ambiguous segments, etc. Lippi-Green endorses Rubin’s interpretation that listeners will perceive an accent even when one is not present but adds to this the notion that the listener is ultimately culpable for the resulting communication breakdown. In other words, negative social bias leads listeners to choose to expend less energy decoding the acoustic signal, resolving ambiguous segments, etc.

Listeners in the Asian face condition are, according to Lippi-Green, shirking their portion of the communicative burden. Following Rubin’s own interpretation, Lippi-Green claims Rubin’s findings indicate that “preconceptions and fear are strong enough motivators to cause students to construct imaginary accents, and fictional communicative breakdowns.” (Lippi-Green, 1997, p. 128). The assumption underlying these claims is that Rubin’s results indicate reduced attention on the part of Asian face condition listeners. Due to racial bias these listeners are simply not attending to the acoustic signal as closely as those in the Caucasian face condition. Given Rubin’s task, though, lack of attention must be inferred from the outcomes rather than observed. Whether listeners actually show reduced attention is an empirical question and one that eye-tracking is particularly well suited to investigating. I will return to this question first in Chapter II when evaluating the extent to which viewing an Asian face can *enhance* perception when hearing a Chinese-accented voice and again in Chapter IV when investigating the time course of listeners’ use of socioindexical information.

1.2 Speech perception research and the neglect of social factors

It is customary in the literature on socioindexical perception to note that social factors have traditionally been neglected by speech perception research. This section provides a brief history of the methodological and theoretical assumptions that may have led to this apparent neglect. This is not intended to defend or exonerate the field's long minimization of socioindexical influences, but to shed some light on the motivation for it.

The central goal of speech perception research is to understand how listeners classify segments of continuous sensory input into discrete mental categories. For nearly 70 years, researchers have sought to understand the system listeners use to map objectively highly variable acoustic signals onto often surprisingly uniform subjective linguistic percepts –a many-to-one mapping. Obviously it is necessary to proceed from some assumptions about the nature of both what constitutes the sensory input in this mapping and what form the mental categories take.

Both speech production and perception make heavy use of acoustic signals, although this emphasis on sound to the exclusion of other sensory input is now understood to be incomplete (c.f. McGurk and MacDonald, 1976; Green et al., 1991; Gick and Derrick, 2009). To understand linguistic mental representations, early researchers turned to the then-dominant structuralist linguistics and its emphasis on sound systems composed of minimally contrastive phonemes (Liberman, 1996; Jusczyk and Luce, 2002). The assumption, among even the most influential researchers in the field, was of a "beads on a string[1]" analogy in which speech perception is the process

[1] This phrase is often attributed to Bloomfield (1933). While Bloomfield clearly supports the notion of speech as a series of individually distinct phonemes, he does not use this phrase.

of using acoustic cues to segment and classify a series of distinct phonetic segments.

Harris (1951, p. 25) describes speech as “a succession of segmental elements, each representing some feature of a unique speech sound.” The methodological mission of the structuralist field linguist was to dissect utterances into these segments, identify the contrastive features of the phonemes of the language and to describe how these phonemes are combined to create distinct morphemes and how these morphemes might, in turn, be combined into larger constructions such as words or sentences.

At the lowest levels, speech perception based on these assumptions can be straightforwardly framed as a psychophysical question: the task of the speech perception researcher becomes establishing the physical limits of the human capacity to distinguish segments of an auditory speech signal, cataloging the particular cues available at the segmental level to uniquely identify particular phonemes that are then available to be used contrastively by some higher-level cognitive system to retrieve words and their meanings.

Of course, strictly psychophysical research might concern itself with identifying the highest and lowest frequencies listeners can detect; the smallest distinguishable frequencies, or just noticeable differences, the human auditory system can recognize; minimum and maximum amplitudes and the dynamic ranges useful in speech. Researchers have long appreciated that speech perception is not isomorphic to psychophysical perception (Abramson and Lisker, 1970; Miyawaki et al., 1975; Babel and Johnson, 2010) (although cf. Appelbaum, 1999). However, the way the field has evolved and investigated the central questions suggests a strongly psychophysical, or more accurately psychoacoustic, pedigree. Much of the focus is on the limits of perception, the influence of the auditory system on speech and careful, rigorous control of experimental methodologies (Repp, 1987).

Three of the key problems[2] that quickly emerge within this research agenda are:

[2] Not all speech perception researchers have agreed on only these three problems; Klatt (1979), for example, adds five more.

1. **Invariance** What are the invariant cues that allow listeners to discriminate particular contrastive segments of speech?

2. **Segmentation** What cues or mechanisms do listeners use to segment acoustic input onto speech sounds that might map onto linguistic objects such as phonemes, morphemes and possibly larger objects?

3. **Talker Variability** How do listeners cope with variation in the speech signal resulting from variability both within and across speakers?

As noted by Jusczyk and Luce (2002), the segmentation and invariance problems emerge largely from the structuralists' conceptualization of hierarchical linguistic structure and the bottom-up mapping of unique segments onto phonemes, phonemes onto morphemes and morphemes onto larger constructions. However, these problems were also encouraged and made intractable by the conceptualization of speech perception as a psychoacoustic phenomenon. If speech perception progresses strictly bottom-up from the raw acoustic signal through a series of intermediary, increasingly abstract, classifications to eventual semantic knowledge and if the researchers' goal is to identify the physical capabilities and limitations of this perceptual system, then the lack of invariance in the acoustic signal really is an insuperable problem. Furthermore, the introduction of, or inability to control for, listeners' cognitive bias –be it in the form of linguistic experience or socioindexical expectation– is methodologically troubling. Listeners' tendency to identify members of a gradient continuum as existing words rather than non-words, for example, will seem to mask and distort the researcher's view of the underlying perceptual system like an ichthyologist trying to study fish without being able to remove them from the water. Even by the late 1960's, though, it was clear to speech perception researchers that linguistic experience does play an important role and that speech perception is not isomorphic to psychoacoustic perception.

1.3 Talker Variability, Normalization and Exemplar Models of Speech Perception

It isn't as though, even in the earliest work, speech perception researchers were unaware of the systematic variability introduced by speaker differences. Peterson and Barney (1952), to name only one classic example, make it clear that there is tremendous variation in the acoustic signal presented to listeners. Their results demonstrate acoustic overlap on the F_1/F_2 dimensions among linguistic categories across age and gender categories. This apparent overlap in category membership is one example of many to pose the vexing problem of the talker variability problem: how listeners derive robustly accurate percepts from highly variable input. The classic approach to data like those in Peterson & Barney has been to hypothesize normalization processes. These processes may be extrinsic to individual speech sounds, in which case a speaker's formant values for different vowels are analyzed in relation to one another (e.g. Joos, 1948; Potter and Steinberg, 1950; Ladefoged and Broadbent, 1957). In this case, it is the structural relationships within vowel systems that are invariant cues to category membership. The listener accurately categorizes incoming segmental information by shifting the entire vowel space for a given talker. In an alternative view, normalization is intrinsic to the speech sounds themselves. In intrinsic normalization, category memberships are assigned by, for example, identifying the ratios between formant frequencies in the input segment (Miller, 1989) or by using a speaker's maximum and minimum formant values to establish the bounds of a normalized formant space into which the entire vowel system can be mapped –thus making vowel systems directly comparable across talkers (Gerstman, 1968). Adank et al. (2004) provide a brief overview and empirical comparison of a large number of normalization procedures.

In each of these approaches, the acoustic signal undergoes a kind of post-processing or re-analysis to strip predictable variability from the signal and form a percept on the basis of the resulting normalized signal. Though by no means the first to do so, Johnson (1997) offered an approach in which acoustic signals are not normalized at all but stored in their entirety and used to activate previously stored episodic traces of speech.

By this activation of exemplars, the word /bɪt/ spoken by a child is unlikely to be interpreted as an adult woman's /bit/ because, in spite of potentially identical F_1 formant frequencies, it excites and is categorized in terms of similarity to stored episodic traces of child speech rather than stored memories that have been previously labeled adult and female. A crucial component of exemplar recognition, therefore, is the ability for listener beliefs about speaker identity to activate socially-indexed features and thus preemptively promote or boost otherwise less stereotypical exemplars (Johnson, 2006). Exemplar theories of speech perception not only assume a great deal of variability in the formation of the lexicon, they also explicitly assume a mental lexicon in which entries store, and are accessed using, the patterns of systematic variability that are methodologically and theoretically daunting to the Motor Theory, Quantal Theory, and other theories of speech perception in which the listeners' lexicon is indexed using only idealized, abstract constellations of gestural information, acoustic landmarks or distinctive features stripped of specific socioindexical (non-linguistic) information. Furthermore, exemplar models of speech perception allow the listener access to fine-grained phonetic detail even after classification has taken place.

Direct Realism posits that perceivers have access to the richly-detailed sensory information resulting from a speech event (Fowler, 1986). Perceivers, in this model, directly apprehend the vocal tract gestures involved in the production of the utterance being perceived with no mediating indirect representations. Listeners' perceptual systems are tuned by experience to perceive "higher-order invariants available in the flow of stimulus information" (Best, 1995, p. 175). There is room for socioindexical effects in a direct realist paradigm. Indeed, a central tenet of the direct realist theory of speech perception is that speech perception must be understood within its ecological niche as one part of the production/perception loop of human communication. Direct Realism is, in principle, entirely consistent with investigation of the influence of socioindexical expectation on speech perception.

However, Exemplar models' emphasis on variability and listener knowledge of social categories has inspired a spate of investigation into the relationship of social knowledge and speech perception in a way that Direct Realism has not. Unsurprisingly, this work has, in turn, generally been interpreted as support for an exemplar model of speech perception. Hawkins (2003), noting the complex multi-modal nature of the proximal speech stimulus, argued for a general shift toward exemplar representations in speech perception research. The *Journal of Phonetics* in 2006 devoted an entire issue to investigations of this question from an exemplar perspective in which Hay et al. (2006b), Johnson (2006), Pierrehumbert (2006), Foulkes and Docherty (2006), and others elaborate on the implementation of sociolinguistic knowledge in an exemplar model. Exemplar theoretic models have even inspired the search for socioindexical effects in perception of syntactic structures (Squires, 2011).

Accessing the lexical entry for a word in an exemplar lexicon consists of calculating, for example, the Euclidean distance between what the listener is hearing and the stored episodic traces of every word he or she has ever consciously heard and attended to (that has not faded from memory). What features actually comprise the probe and the stored exemplars is surprisingly unclear in the literature. Hintzman (1986) describes modality-specific sensory features (in the case of speech: pitch, amplitude, frequency, etc. as transduced by the ear) but also emotions and more abstract features (Hintzman lists: "before, same as, greater than, has as parts"). For speech perception, Johnson (1997) models a vowel disambiguation task using 5 features: fundamental frequency; first, second and third formant frequencies; and vowel duration – implying that a phoneme-level segmentation process has already operated upon acoustic input prior to storage or retrieval of the phonemes of a word. Pierrehumbert (2001) assumes feature representations of whole words are stored in the exemplar space and phonemes are the intersection of these traces (accounting for listener judgments that /bɛt/ and /bɛnt/ share the same vowel (distant neighbors of an /ɛ/ exemplar neighborhood (or 'cloud')).

This general approach was not entirely new to linguistics in 1997. Repp (1987) describes a system in which the listener selects a lexical entry by calculating the perceptual distance between the incoming phonetic information and a permanent store of possible mental representations of phonetic alternatives which he describes as prototypes. It has been suggested (e.g. Hawkins, 2003; Port, 2008) that Klatt (1979) describes an early precursor to exemplar models. Klatt describes the Lexical Access from Spectra (LAFS) model in which lexical entries are stored as a network of phonemic representations. Each phoneme pairing has a stored diphone spectral representation in one of approximately 300 spectral templates. The spectral templates represent critical band diphone spectra for all phonotactically-possible phoneme pairs in the lexicon where each critical band represents a frequency region corresponding to a constant length of the ear's basilar membrane. Prior to receiving input, each lexical entry is precompiled into a finite state automaton-like network of these diphone templates and lexical access proceeds by decoding the path through this network that best matches the acoustic input. It could be argued that this Klatt model implicitly incorporates an intrinsic normalization procedure by mapping raw acoustic input onto critical bands, but as these bands will already have been imposed by the transduction of the peripheral auditory system it is difficult to see how one could object to positing that mental representations are composed of these bands.

What Repp (1987) and Klatt (1979) lack, and Johnson (1997) provides, is an explicit way of storing talker-specific or social category-specific knowledge and incorporating that knowledge into subsequent lexical access. This is the key feature that makes exemplar models compelling to those interested in modeling talker-specific and social category effects in speech perception. Another feature of exemplar models that makes them attractive to socioindexical perception research is the retention of fine phonetic detail that is not obviously relevant to the linguistic or referential decoding of the speech stream. Retaining this additional phonetic detail makes it possible to argue that listeners might use this detail to access social category and individual identity representations in a socioindexical lexicon (Munson, 2010). Exemplar models' response to the lack of invariance problem and the talker variability problem is to embrace variability and account for variability directly in listeners' mental representations by supposing that listeners are sensitive to even extremely subtle patterns of variation in speech; that is, by accounting for variability in grammar.

I will now turn to a review of some of the key findings in socioindexical and related speech perception research.

1.4 Identical acoustic input: variable perception

In the simplest terms, socioindexical perception research suggests that a listener's mapping of a given acoustic signal onto mental representations can be made to vary by holding the acoustic signal itself constant but varying the listener's beliefs about the identity of the speaker. This turns knowledge of variability on its head by hypothesizing that not only is lexical access robust to variation in the signal because the listener has knowledge of systematic variability, but listeners can have their most-probable mental representations shifted by expectation of socially-indexed variation.

This lack of correspondence between objectively measurable acoustic stimulus and reported percept should not be surprising as it is the rule rather than an exception. Ladefoged and Broadbent (1957), for example, found that listeners' percepts of target words at the end of a carrier phrase could be shifted by altering the vowel space in a carrier phrase through a process that Ladefoged and Broadbent hypothesized to be extrinsic vowel normalization.

Even in a phenomenon as familiar to linguists as allophonic variation we encounter a case in which identical acoustic information is reliably and robustly mapped onto different mental representations depending on, for example, the location of the acoustic information within a syllable. Therefore the short-lag voice onset time of the second phonetic segment in [spɪt] will be perceived as a /p/ in that context but if one edits the waveform to remove the initial fricative material listeners will robustly and consistently perceive the identical acoustic input as /b/ in a new percept of [bɪt]. Similarly, Coetzee and McGowan (2008) found that listeners are so keenly sensitive to allophonic variation when perceiving syllable boundaries that they perceive an illusory schwa through perceptual epenthesis; acoustic input consistent with [sp^{h}ika], where the long VOT [p] allophone was spliced from a word-initial position, was nevertheless perceived as [sə.p^{h}i.ka]. Dupoux et al. (1999) have documented similar perceptual epenthesis by Japanese listeners.

The perceptual mapping of acoustic input can be altered by its phonetic context regardless of whether that context precedes or follows the target segment. Miller and Liberman (1979) manipulated formant transition duration to create a continuum of stimuli from [ba] to [wa]. They found that listeners' percept of word-initial [b] and [w] could be shifted by altering the length of the steady state vowel formants in the synthesized syllable. More of the continuum was identified as [wa] when the overall syllable duration was short, with more [ba] responses as syllable length increased.

McGurk and MacDonald (1976), in which misleading visual information about oral stops in CV syllables is shown to override otherwise clear acoustic information ("the McGurk effect"), is probably the most famous example of the auditory signal being supplemented or even obliterated by other context during low level speech perception. Vatikiotis-Bateson et al. (1998) tracked listener eye movements to a video of a speaker's face and found that listeners fixated on the speaker's mouth more as signal to noise ratio increased. Far earlier, though, Liberman et al. (1952) found that an identical 1440Hz burst is perceived as a [p] before the vowel [i] but as a [k] before the vowel [a] suggesting, at least from one perspective, that listener expectations about articulation can alter the perception of even non-speech acoustic information in much the same way as the McGurk effect. Similarly, Lisker and Abramson (1964) and Lisker (1967) found that listeners perceive members of a VOT continuum categorically with a sharp discrimination boundary whose location is dictated, at least in part, by one's native phonology. Ganong (1980) found an analogous effect for words; listeners hearing a continuum from one of seven word to nonword or nonword to word continua such as *tash* to *task* tended to perceive the stimuli as the real word.

Mann (1980) found that listeners' labeling of consonant continua can be shifted by manipulating the identity of nearby segments in the stimulus. The consonant [l] typically has a high frequency F_3 offset while, by contrast, [ɹ] tends to have a low frequency F_3 offset. [d] typically has a high frequency F_3 onset while, again by contrast, [g] has a low frequency F_3 onset. Due to coarticulation, we would expect a naturally-occurring [d] to have a lowered F_3 onset after an [ɹ] and we would expect a naturally-occurring [g] to have a higher F_3 onset following an [l].

Mann found that listeners presented with members of a synthesized [da]-[ga] continuum varying in F_3 onset will perceive more members of the continuum as (typically low F_3 onset) [ga] when they are preceded by (high frequency F_3 offset) [al] and more members of the continuum as (typically high F_3 onset) [da] when preceded by (low frequency F_3 offset) [ar]. This phenomenon is generally referred to as 'compensation for coarticulation[3]'.

Mann interpreted this result to indicate that listeners anticipate a higher F_3 fre-

[3] I believe this effect can more accurately be described as 'expectation of coarticulation'. The term 'compensation' implies that coarticulation is primarily deleterious to the signal and is firmly rooted in the 'beads on a string' analogy.

quency following an [l] so the larger number of [ga] labels in an [al] context indicate that listeners attribute the higher F_3 in those members of the [da]-[ga] continuum to coarticulation. Others have argued for a spectral contrast interpretation, but what this result minimally indicates is that listeners' percepts of identical acoustic information can be shifted between [da] and [ga] simply by altering the phonetic context.

Recently, Gick and Derrick (2009) have even found that, like the visual input in the McGurk effect, puffs of air can override the auditory signal and influence listeners' perception of VOT. As with the allophonic variation already described, listeners hear identical acoustic information as [b] in the absence of a puff of air on the hand or neck and as [p] when the puff is present. Responses on control trials, when the tactile input was withheld, show no such pattern. These results suggest that, as with visual and auditory information, listeners integrate somatosensory input when arriving at a speech percept and this somatosensory input can shift the perception of identical acoustic information.

All of these disparate findings point to a single conclusion. The mapping of acoustic stimulus to mental representation is surprisingly malleable. Classification of the acoustic stimulus can be shifted by visual information about articulation, somatosensory information hinting at aerodynamics, coarticulation, speech rate, native language(s), lexical status, and more. Listeners will incorporate any information available to them when perceiving speech. In all of these cases, listener beliefs or expectations lead them to report a perceptual experience that is not uniquely determined by the acoustic stimulus.

Most of the research described in this section has in common the observation that perception of a particular segment can be altered by manipulating aspects of the perceptual input other than that segment's acoustic information. In all cases so far, though, this additional sensory information, be it auditory, visual, or somatosensory, has related directly to the linguistic aspects of the acoustic signal. Socioindexical perception designs manipulate sensory input that would traditionally have been assumed to be irrelevant to the linguistic aspects of the speech stream. The key difference, then, between the traditional speech perception research described so far and socioindexical speech perception research is the relaxation of this assumption of irrelevance. The goal of understanding how listeners classify segments of continuous sensory input into discrete mental categories remains the same.

1.5 Socioindexical speech perception

A claim in the socioindexical speech perception literature is that expectation of socially-indexed variability affects speech perception in much the same way, and at the same level of processing of phonetic information, as the research discussed in the previous section. I will argue in this dissertation that this assessment may be premature given the kinds of evidence available.

Strand (2000) investigated the influence of gender stereotypes on low-level speech processing and found that listeners recognize words more slowly when the pitch of the speaker's voice is atypical of his or her gender. She concluded that stereotypes can speed processing –stereotypical features are processed more quickly than non-stereotypical features or features which defy stereotype. Green et al. (1991), however, created McGurk style stimuli in which the normal McGurk mismatch was introduced alongside a gender mismatch – female faces with male voices and male voices with female faces. They do not provide Strand's thorough, compelling assessment of gender typicality, but they do find that the McGurk effect is robust to gender mismatches. At a minimum I believe this suggests that the integration of socioindexical knowledge during speech perception is somewhat more interesting than a simple interpretation in which expectations are incorporated identically to manipulation of referential aspects of the speech stream.

McLennan and Luce (2005) found that varying indexical factors could slow recognition of difficult words. In an eye tracking study, Creel et al. (2008) found that lexical competition, indexed by fixations to the target word, between such near neighbors as *sheet* and *sheep* is diminished when the words are spoken by different talkers. The identity of the speaker, in these experiments, influences word recognition and lexical competition effects that have been traditionally investigated strictly in terms of linguistic aspects of the speech signal.

As suggested by the Rubin (1992) study, though, it also seems to be the case that the mere *expectation* of social difference can similarly alter listener judgments. Hay et al. (2006b) investigate the in-progress merger of the diphthongs /iə/ and /eə/ in New Zealand English and find that, among other factors, ostensible age and social class of the speaker biased the rate at which identical auditory stimuli were perceived as, for example, *ear* or *air*. This experiment uses a matched guise variation of a two alternative forced-choice reaction time experiment in which the acoustic stimuli are held constant and digital images of purported speakers are used to manipulate listeners' perception of the purported speaker's age or social class. Identifying an isolated word one has heard is a somewhat unnatural linguistic task, but the listener responses are immediate and relating their accuracy and reaction time to the processing of the acoustic input in light of the face manipulation is similar to methodology employed in more traditional speech perception research.

Staum Casasanto (2009a) investigated listeners' expectations[4] of African American English (AAE) and the role of those expectations during speech perception. In one example of her experiments, listeners hear a phrase such as "The mass probably lasted..." while looking at a white or black face whom they believed to be the speaker. The participants are then presented with a continuation of the sentence and asked whether the continuation makes sense. The continuation 'through the storm'

[4] Note: *expectation* is my term; Staum Casasanto refers instead to *knowledge*.

would be consistent with AAE *t/d* deletion in (*mast*) and ‘an hour on Sunday’ would be consistent with non *t/d*-deleted (*mass*). Participants were faster to identify the *t/d* deletion-consistent continuation when they believed the speaker to be African American and slower to identify the non-*t/d* deletion-consistent continuation. Staum Casasanto argues that this sensitivity is derived from listeners’ knowledge of the statistical regularities observable in AAE and the influence of that knowledge on the processing of phonetic detail in the target word.

It is clear that Staum Casasanto’s listeners are showing different behavioral outputs on a sentence completion task given the matched guise manipulation. It is not at all clear from the task that this behavior is due, as she interprets it, to altered bottom-up processing of phonetic detail. It could well be, with such a late measure, that socioindexical expectation is exerting a top-down influence and allowing the listener to select a socially, as well as phonetically, appropriate lexical item. It is similarly not clear that this result is evidence for an exemplar model of speech perception. Exemplars certainly offer a compelling means of associating social category labels with lexical items, but performance on this sentence-completion task does not seem capable of differentiating one possible storage hypothesis from another.

1.6 Dissertation Overview

To address the goals laid out in this introductory chapter, I investigate experienced and inexperienced English speaking listeners’ perception of Chinese-accented speech. ‘Experienced’ listeners are heritage speakers of Mandarin Chinese with extensive experience listening to Chinese-accented English but limited competency speaking the language–a population of listeners Au et al. (2002) refer to as ‘childhood overhearers’. ‘Inexperienced’ listeners are native English speakers who report little or no experience with any dialect of Chinese or with Chinese-accented English.

The choice of social category to manipulate was a difficult one. Chinese-accented speech was selected to satisfy two main criteria. First, it was important to attempt to control for and investigate the influence of real world knowledge that listeners bring with them into the laboratory. I anticipated that the selection of Chinese-accented speech would simplify the process of identifying and recruiting populations of experienced and inexperienced listeners. Secondly, I wished to be able to build upon, using a more on-line task, the first experiment of Rubin (1992) to evaluate the negative bias hypothesis offered by Rubin and the yet stronger claims of Lippi-Green (1997).

Chapter II addresses the difficult problem of quantifying listener knowledge, experience, and stereotypes under laboratory conditions. I report the results of an experiment in which listeners were asked to complete an authentic Chinese-accented English identification task. Populations of both inexperienced and experienced listeners participated. Performance is evaluated, compared, and discussed alongside the participants' self-identification on a language history survey instrument. Experienced listeners are both more accurate at identifying an authentic accent, and less likely to be convinced by an imitated Chinese accent than inexperienced listeners. I discuss the implications of these findings for exemplar models –including the difficulty of interpreting this unnatural task as either evidence in support of or against the predictions of an exemplar model of speech perception.

Chapter III reports the results of a sentence-in-noise transcription task using the matched guise technique common in sociolinguistic work on the perception of socioindexical features (for a detailed review of the literature on matched guise see Campbell-Kibbler, 2005, Chapter 3). The same experienced and inexperienced populations of listeners as those reported in Chapter II participated in this task. Listeners in a between-subjects design were presented with an image of the face of their purported speaker and asked to orthographically transcribe Chinese-accented sentences embedded in multitalker babble. Overall, transcription of sentence-final keywords was significantly more accurate when keywords were highly predictable and the presented face was Asian than when the face presented was Caucasian. This finding nears significance for the small population of experienced listeners and is statistically significant for *inexperienced* listeners.

This outcome with inexperienced listeners suggests the existence of higher level stereotypical phonological abstractions derived from some source of knowledge other than extensive personal experience with a particular variety. These stereotypical features are, nevertheless, believed by the listener to index that variety. Listeners with little or no experience listening to native speakers of Chinese-accented English can usefully employ socioindexical knowledge to enhance perception of speech in noise.

In Chapter IV, I report the results of an eye-tracking investigation of the time course of listeners' use of socioindexical knowledge. Participants in a visual world eye-tracking paradigm were presented with a Standard American English voice and the image of either an Asian or Caucasian face of the purported speaker. Surprisingly, listeners appear to trend somewhat *faster*, though not significantly so, to fixate the correct alternative in a two-alternative forced choice when the face is Asian. These listeners also fixated this image for a significantly longer period than listeners in the Caucasian face condition. This result is contrary to experimenter expectations and not obviously in keeping with exemplar theories of speech perception in which socioindexical expectations pre-activate clouds of socially-labeled exemplars.

Finally, in Chapter V I briefly offer conclusions resulting from the dissertation as a whole.

CHAPTER II

Experienced and Inexperienced Listeners: yes/no task

The goal of this chapter is to quantify the ability of listeners to accurately identify authentic Chinese-accented English. The performance of self-identified inexperienced and experienced populations of listeners will be tested and compared. It will then be possible to directly test the frequent attribution of socioindexical effects in speech perception to stored episodic traces of linguistic experience labeled with social knowledge.

An additional, methodological goal is to explore the usefulness of this identification task as a means of directly estimating participants' experience level with authentic Chinese-accented English. Finally, this experiment lays the groundwork for a larger project, outside the scope of this dissertation, exploring the acoustic correlates of 'Chinese-ness' for both experienced and inexperienced listeners.

2.1 Ideology & Expectations

Quantifying listener expectations about, and experience with, language is a daunting task but one that is frequently necessary in laboratory work. This problem is by no means unique to questions of socioindexical experience nor indeed even to speech perception. In diverse linguistic or psychological experiments it is often necessary to assess the frequency of a particular phoneme, word, transitional character probability, n-gram frequency, etc.

The normal practice is to calculate these frequencies from an established corpus such as Kucera and Francis (1967) or Baayen et al. (1993)[1]. Although both corpora are quite dated and drawn from a mixture of print and spoken sources that are unlikely to represent the statistical patterns experienced by modern participants (Balota et al., 2007), these data provide an expedient and, more importantly, standardized surrogate for listener experience with linguistic forms. In the following section I describe the particular importance of quantifying not only listener experience but also listeners' language ideologies in socioindexical perception research.

In this chapter I explicitly draw on the concept, from linguistic anthropology, of a *language ideology*. 'Ideology' is, in many respects, an unfortunate term for a useful idea. Unfortunate because the word itself indexes a kind of self-indulgent conspicuous intellectualism. Ideologies in this sense are a system of ideas or beliefs through which listeners and speakers link linguistic forms to social groups, people and even events or activities (Irvine and Gal, 2000, p. 25). This linkage is bi-directional so that listeners come to associate particular linguistic features with particular social groups (e.g. metathesized 'ask' in American English). Listeners also associate social groups with sets of linguistic features (e.g. native speakers of English asked to imitate a regional dialect or foreign accent will be remarkably consistent in the features they choose to perform to index that language variety). Irvine and Gal describe how for the Nguni the click consonants of the nearby Khoisan languages indexed 'conspicuous foreignness'. This ideology led to the Nguni adoption of clicks to replace native phonemes to create and maximize differentiation from normal speech for an avoidance register.

[1] frequencies from the COBUILD corpus

I believe this concept of ideology is useful not only for understanding and describing the behavior of speakers, but potentially for understanding the expectations listeners bring to the tasks of understanding language and of perceiving speech.

2.2 Quantifying Listeners' Experience

Work in socioindexical speech perception critically relies on an understanding of the identity of the listeners. The need to quantify listener experience is therefore likely greater than in other linguistic and psycholinguistic experimentation. Here we have all of the same questions of frequency and patterning of linguistic forms but with the added recognition that these forms will differ systematically by listener and social context.

It is of course extremely useful to conceive of identity as a monolithic, constant feature of an individual, but this is also massive simplification. In reality, it would seem that identity is a dynamic, context-sensitive construct in which interlocutors manipulate and interpret indexical forms to define their roles in a particular interaction (Bucholtz and Hall, 2010). Irvine (1989) refers to "a diversity on the linguistic plane that indexes a social diversity" and recent work in sociolinguistics and linguistic anthropology has demonstrated that speakers and listeners are aware of, and exercise situational control over, these diversities. Depending on context, speakers will invoke different constellations of indexical linguistic forms –different registers– to convey the same denotational or referential meaning (Silverstein, 2003). In other words, not only does one speak differently in a job interview than one speaks in casual conversation, but listeners are aware of and expect this use of appropriate registers.

A. Babel (2010) reports the use of Spanish/Quechua contact features among speakers in one Andean village. Given local ideologies that associate Quechua use with informal, rural speech, it is unsurprising that Quechua-influenced contact features in this variety of Andean Spanish are more commonly used in informal conversation than in interview or meeting contexts. However, these features are also invoked in more formal contexts as indices of the speaker's authenticity, to create intimacy, to mark one's affiliation with a particular political group and sometimes several of these social meanings simultaneously. Individual linguistic forms may be linked to particular social meanings, but both the linkage and the meaning are highly context-dependent.

For the experimentalist, then, it is worth bearing in mind that performance on a task intended to quantify listener experience will be shaped by the formal, experimental context, by listener ideologies about the details of the language being used and by listener ideologies about the purported speaker.

For the purposes of this dissertation, which investigates inexperienced and experienced listeners' use of visual socioindexical cues when processing acoustic input, it would be ideal to have a range of background information about each participant that it is difficult to conceive of collecting either behaviorally or through self report. This includes such variables as the frequency and intensity of interaction with Asian interlocutors, the probability of an Asian face accompanying a non-native English accent in the listener's experience, the distribution of facial features and L1 languages and their combination in the listeners' experience with speakers, how the listener construes the phonetic features under investigation to construe meaning, how the listener self-identifies linguistically, etc. Every aspect of stimulus presentation is potentially open to influence from listener experience and ideologies.

This depth and breadth of understanding is difficult, if not impossible, to achieve in the laboratory. The standard solution is to ask participants to complete a language history survey –either in the laboratory or when registering for a subject pool. Survey instruments vary, but they generally request such information as the participant's native language(s), languages spoken at home, languages studied, places the participant has lived, etc.

One promising method of assessing participants' unconscious social biases that has recently been used in speech perception research is the Implicit Association Test (IAT) (Greenwald et al., 1998). In an IAT instrument, the experimenter creates word lists of equal length in four categories: a set of words from one end of an evaluative continuum (e.g. ‘loyal’), a set of words from the other end of that continuum (e.g. ‘disloyal’), a set of words associated with one social group (e.g. catholic) and finally a set of words associated with another social group (e.g. protestant). Through a series of blocked presentations and responses the participant is asked to categorize the first 2 sets of words as loyal, with one button, or disloyal, with another, and categorize the second two sets of words as ‘catholic’ or ‘protestant’ in the same manner. In a final set of presentations, participants are asked to categorize the half of the words while loyal/catholic are linked to one button and disloyal/protestant are linked to the other and then the pairings are reversed (disloyal/catholic and loyal/protestant) for the second half of the words. Participants who would be unwilling to self-report social biases, or who may even be unaware of unconscious social bias, will show an increased response time when a word they associate with one attribute is paired with the label for the opposite social group. A member of the Irish Republican Army, for example, might show slower reaction times when ‘loyal’ words are linked with ‘protestant’ whereas a member of the Orangemen might show the same slowed reaction time when a ‘loyal’ word is linked with ‘catholic’.

As the example labels I have chosen imply, the test depends on a complex set of contextually and culturally specific ideological linkages with each set of words. The test also depends on the participant activating the intended indexical meaning of each word. For example, ‘loyalist’ in the context of Northern Ireland is a person who is strongly in favor of Northern Ireland’s membership in the United Kingdom as opposed to the unification of that portion of the island with the Republic of Ireland, which is not part of the UK. An experimenter unaware of this culturally specific meaning of ‘loyal’ may create dichotomies that suggest precisely the wrong implicit association (or mask that association).

It is surprising that something so apparently tenuous can provide consistent results, but the IAT has seen wide usage in social psychology. In linguistics, M. Babel (2009) used sets of 'white names' and 'black names' and a 'good words'/'bad words' dichotomy to detect implicit racial biases in speakers participating in an accommodation experiment. Among many other influential factors, Babel found that participants with a pro-black bias on the IAT were more likely to have their vowels converge toward those of a black talker.

Christy (2010) also used the IAT task to test the influence of implicit white/black racial bias on listeners' perception of a phonetic feature reported to index African American English. Christy reinforced the results of the IAT with a traditional survey instrument. Participants also completed a language experience survey in which they rated their own time spent with children each week on a 10 point scale from 'none' to 'extremely frequently'. Fascinatingly, listeners with little or no experience listening to children's speech showed a strong correlation between their IAT-measured racial bias and ratings of speech. On the other hand, listeners with experience interacting with children showed no such correlation.

Staum Casasanto (2009a) addresses the question of experience, which she refers to as knowledge, and language ideologies, which she frames as stereotype, in a written survey. Participants read 24 sentences: 12 containing an apostrophe to indicate word-final t/d deletion (a feature consistent with consonant simplification in AAE) and 12 sentences containing other 'nonstandard' usages not known to index AAE. Participants associated the t/d deleted sentences with a photograph of a purported African American speaker 60% of the time while the other nonstandard features were associated with the African American speaker only 51% of the time (Staum Casasanto, 2009a, p. 83). It is unclear whether this task truly differentiates experience from stereotype, though, as the other nonstandard features used are also stereotypes which strongly index other social groups (e.g. the ‘cawfee’ stereotypical of Long Island and the ‘needs washed’ construction of the Midland dialect).

The approach to listener experience quantification that I take here adapts a task from forensic phonetics (Neuhauser and Simpson, 2007) and is essentially an attempt to assess participants’ ability to correctly identify an authentic Chinese accent from a set of distracter accents. It must be acknowledged that the generalization ‘Chinese accent’ is so broad as to be almost comical[2]. The so-called ‘dialects’ of Chinese comprise 6 separate language phyla: Sino-Tibetan, Austro-Tai, Austronesian, Altaic, Austro-Asiatic and Indo-European. Many of these dialects are not mutually intelligible, with listener subjective ratings of mutual intelligibility closely matching performance on cross-dialectal semantic classification and speech-in-noise perception tasks (Tang and van Heuven, 2009). This suggests that, although Chinese L2 English speakers may all *also* command Mandarin, or Standard, Chinese, Mandarin is itself likely to be an L2 language or spoken with the accent of a regional dialect. Additionally, different non-native English speakers from China have acquired different target Englishes. Until quite recently it has been the norm for Chinese students of English to target RP-accented British English as their normative model for acquisition. Increasingly, though, students target American English or even a contact variety known as ‘China English’ (Qiong, 2004).

The projects described in the remainder of this chapter represent first steps toward a more comprehensive investigation. This work has been useful in informing and shaping the experiments in the following chapters but is not complete. In the discussion for this chapter I will describe the next steps for this line of research, as I now understand them, and the lessons learned so far.

[2] A problem I encountered frequently that I have yet to understand completely was participant assumptions that by ‘Mandarin Chinese’ I actually meant Cantonese. Several of the pilot listeners for experiment 1 and one of the actors explicitly hired to imitate Mandarin Chinese-accented English reported expecting that Cantonese was the object of investigation. One listener even explained that he assumed that I, as a Caucasian American, was not aware of the distinction between Cantonese and the requested Mandarin Chinese and must therefore have meant Cantonese.

2.3 Experiment 1: Identifying Authentic Chinese-accented English

Experiment 1 is a *yes/no* accent detection task. The listener is presented with a single stimulus recording per trial and must press one button on a response box if the stimulus sounds like authentic Chinese-accented English and another button if the stimulus sounds like some other form of accented-English. This experiment was designed to detect listeners' ability to correctly detect an authentic Chinese accent among a collection of accents that include authentic Chinese, imitated Chinese, Korean, and other accents.

The primary goal of this experiment was to quantify the extent to which listeners with little or no experience listening to native speakers of a target variety nevertheless use socioindexical knowledge in a systematic way during perception and to compare this performance to that of experienced listeners. Populations of inexperienced and experienced listeners can be identified and their experience quantified using this method. It is then possible to test the frequent attribution of socioindexical effects in speech perception to stored episodic traces of linguistic experience labeled with social knowledge. An additional, methodological goal was to explore the usefulness of this task as a means of directly estimating participants' experience level with authentic Chinese-accented English. Finally, this experiment lays the groundwork for a larger project exploring the acoustic correlates of 'Chinese-ness' for both experienced and inexperienced listeners.

2.4 Methods

2.4.1 Stimuli

Stimulus materials consisted of the eight sentence types listed in table 2.1. These were all English recordings spoken by two native speakers of Mandarin Chinese and one native speaker each of Korean, Turkish and Macedonian all drawn from the Wildcat Corpus (Van Engen et al., 2010). The Wildcat corpus includes individual words, the “Stella” passage from the Speech Accent Archive at George Mason University, the “North Wind and the Sun” from the IPA Handbook, high and low predicability sentences, and unscripted recordings from a map task. The sentence recordings from this experiment were drawn from the scripted passages and sentence recordings. These stimulus recordings were augmented with recordings of two monolingual English speakers performing imitated Chinese accents.

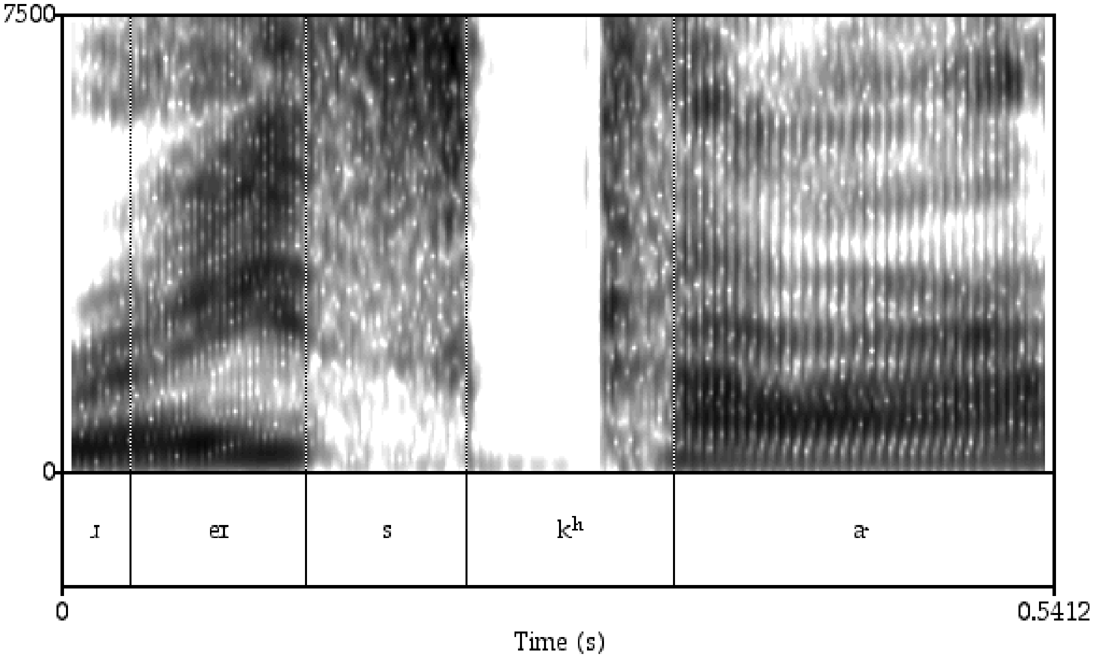

Figure 2.1: **Spectrogram of male authentic Chinese speaker producing *racecar*** Post-vocalic /ɹ/ is clearly absent in the spectrogram.

In general, the accuracy of the imitated Chinese was poor but consistent. Native speakers of midwestern American English from Michigan were asked to read, with an imitated Chinese accent, the same texts recorded by authentic Chinese-accented speakers for the Wildcat Corpus. The voices selected for inclusion in the study imitated the authentic backing of interdental fricatives (/ð/ → [z] and /θ/ → [s]) and the stereotypical feature /ɹ/ → [l] that is rarely, if ever, found in au-

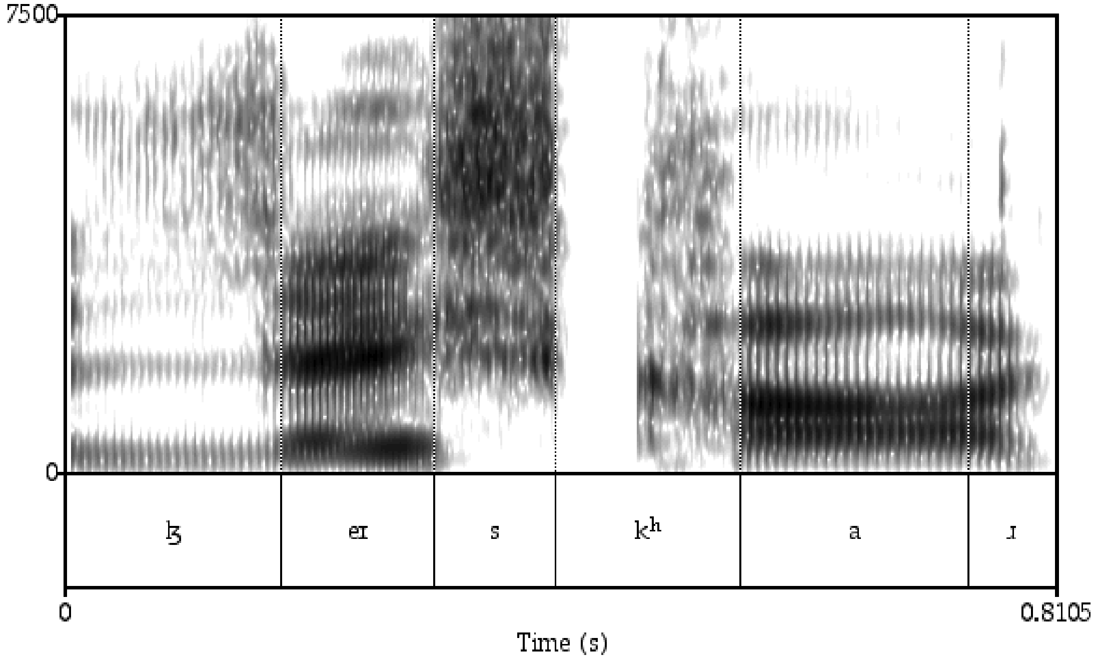

Figure 2.2: **Spectrogram of male imitated Chinese speaker producing *racecar*** In this imitation, initial /ɹ/ has been replaced with a voiced alveolar lateral fricative. Post-vocalic /ɹ/ is clearly visible in the spectrogram.

thentic Chinese-accented speech. Surprisingly, the native American English speakers who produced the imitated Chinese consistently produced post-vocalic /ɹ/ while the authentic Chinese-accented speakers did not.

Figure 2.1 shows a spectrogram of a sample token of authentic Chinese-accented English. This male speaker has produced the word *racecar* as [ɹeɪsk^{h}a˞]. The word-final vowel is rhoticized for nearly its entire duration with no audible consonantal articulation. There is a pitch contour on this syllable similar to the Mandarin Chinese third tone with its characteristic dip and rise.

Figure 2.2 shows a spectrogram of a sample recording of imitated Chinese-accented English. This male speaker has produced the word *racecar* as [ɮeɪsk^{h}aɹ]. This speaker generally replaced /ɹ/ in non-post-vocalic positions with [l] however, in this particular token there is visible and audible frication. The post-vocalic /ɹ/ is clearly visible and audible over the last 71ms of the token and the vowel is audibly rhoticized for

50ms (6 glottal pulses) prior to the consonantal articulation. That the imitated Chinese speakers consistently produced post-vocalic central approximants is initially surprising. The lack of post-vocalic /ɹ/ is a stereotypical feature of Chinese-accented English and one that professional actors in a subsequent study (not reported here) consistently imitated.

I do not have articulatory data for these imitated productions. I also do not have recorded samples of these speakers producing the target sentences in their normal speaking voices. However, very little of this vowel is rhoticized, and this production is typical of the imitated recordings. These speakers speak a rhotic dialect and one might normally expect the final vowel in this word to show extensive evidence of coarticulation with the word-final /ɹ/ (Olive et al., 1993, p. 220). I believe these tokens actually do represent the imitators' attempts at a reduced consonantal gesture.

She made the bed.
Bob wore a watch on his wrist.
Dad talked about the bomb.
I wear my hat on my head.
The color of a lemon is yellow.
A racecar can go very fast.
He looked at the sleeves.
Please call Stella.

Table 2.1: **Sentences used in Experiment 1**

2.4.2 Procedure

Listeners used Apple Macbook Computers (model 4,1; late 2008). Testing with inexperienced listeners took place in an IAC sound-attenuated booth in the University of Michigan Phonetics Lab; stimuli were presented over AKG K271 mkII headphones. Responses were entered via Cedrus RB-620 low-latency response boxes with serial to USB adaptors.

Experienced listeners used the same computers and software as inexperienced listeners. However, these testing sessions took place in the phonology laboratory at the University of California, Berkeley. This is a quiet space dedicated to speech perception experiments, but is not a sound-attenuated booth. AKG k240 headphones and Cedrus RB-730 low-latency usb response boxes replaced the headphones and response boxes used at Michigan.

Stimuli were presented using Superlab stimulus presentation software version 4.0.8. Volume was set at a comfortable listening level. Listeners indicated their responses via button box. Listeners were instructed to press one button if the voice they hear has an authentic Chinese accent and another if the accent is not authentic Chinese. The target sentences were presented on-screen from the onset of the recording playback until the subject submitted a response. Listeners were informed that the voices would include a range of different non-native English accents including Chinese, imitated Chinese, Korean, Turkish and Macedonian. It was not possible to change responses or to hear recordings more than once. Listeners were encouraged to rest after each block and there were enforced breaks at the halfway point.

Based on preliminary results from two pilot experiments, a small effect was anticipated in the data and consequently rather a large number of trials were administered per participant. Participants responded to 8 sentences produced by 7 voices in each of 6 blocks for 336 responses per participant.

All participants in this experiment had just completed the speech-in-noise transcription task reported in Chapter III. No voices or stimuli were repeated from that experiment.

2.4.3 Participants

Eighty-seven undergraduate students participated at one of two experiment sites: the University of Michigan phonetics lab or the University of California, Berkeley phonology lab.

2.4.3.1 Inexperienced Listeners

Fifty-seven undergraduate students from the University of Michigan Introductory Psychology subject pool participated for partial course credit. Participants had no known hearing problems. Five participants were identified for exclusion prior to analysis for reporting experience with Mandarin Chinese –either through language study or, in four cases, for being bilingual or Heritage speakers. These participants will be included in the correspondence analysis but excluded from other statistical and visual data analysis. One participant was excluded for using Facebook and sending text messages on his smartphone. One additional participant was excluded from the data analysis for struggling to remain awake during the experiment and reporting the task was extremely difficult. Three data files were lost due to experimenter error.

2.4.3.2 Experienced Listeners

Identifying a sufficiently large experienced population of Chinese-English listeners at the University of Michigan proved problematic. Heritage speakers with little or no proficiency in Mandarin were selected as a target population early on. This selection was intended to avoid, at one extreme, the complications of interpreting the results of truly bilingual speakers for what is essentially an English language task. At the other extreme, Rubin and Lippi-Green hypothesize a confound for our purposes with native English speakers exposed to Chinese-accented English through native Chinese professors and graduate student instructors. If these monolingual English listeners are refusing to attend to their Chinese-accented instructors then they do not, in fact, represent an experienced population. Experiment 1 suggests that this hypothesis of Rubin and Lippi Green is not correct, but this confound would have made their prediction difficult to refute.

Thirty Heritage Mandarin-speaking undergraduate students from the University of California, Berkeley participated in exchange for an incentive of $15.00 per participant. Two participants were removed prior to any analysis: one L1 speaker of Mandarin who had misunderstood the flier, and a second individual who misrepresented his identity. As with the excluded listeners from the Inexperienced group, these participants will be included in the correspondence analysis but excluded from other statistical and visual data analysis. Time constraints limited the number of participants who could be engaged in the study.

2.4.4 Predictions

The use of imitated Chinese accents was inspired by Neuhauser and Simpson (2007), who found that German monolingual speakers were more likely to identify German imitations of French and American accents in a naming task than they were to correctly discriminate true non-native accents. I hypothesized that native listeners must be drawing on language ideologies concerning foreignness in general and the target non-native accents in particular when making discrimination judgments. If true, this finding would have implications for research in socioindexical speech perception that has appealed to stored episodic traces to explain behavioral results.

It is difficult to imagine a means of differentiating between listener knowledge gained through real communicative experience with a language variety and listener knowledge of linguistic stereotypes (again, in the sense of Labov, 1994) gained through exposure to imitations of that variety or occasional brief exposure in the media. However, the Neuhauser and Simpson (2007) result suggests one possibility. If inexperienced and experienced listeners are drawing on both qualitatively and quantitatively different forms of knowledge when detecting an authentic Chinese accent then they should be differentially drawn to authentic and imitated stimuli.

2.4.5 Results

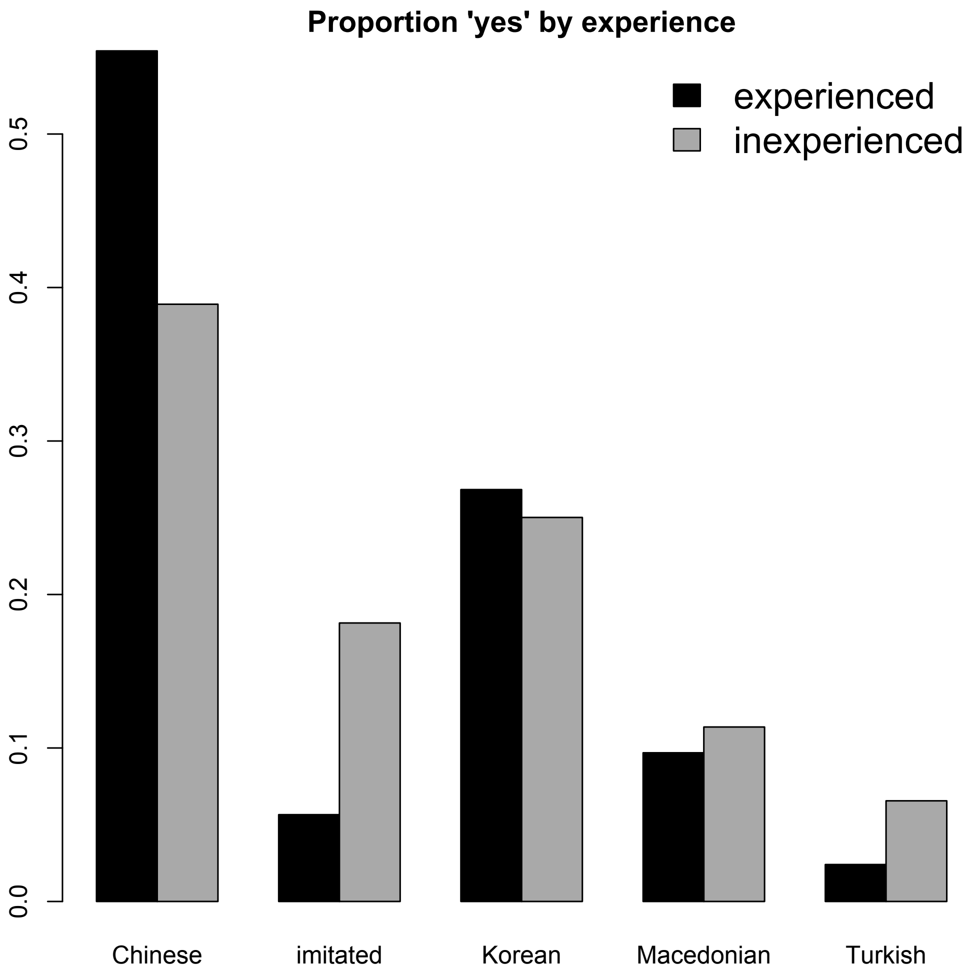

Figure 2.3: **Proportion 'yes' responses by accent and experience.** Bars sum to 1 within experience level.

2.4.5.1 Proportion 'yes' responses

Figure 2.3 shows proportional 'yes' responses to each non-native accent by experience level (bars sum to 1 within each series 'experienced' and 'inexperienced'). Experienced listeners appear to be dramatically more likely to respond 'yes' to an authentic Chinese voice (the only technically 'correct' response) than to any other

(ref. level: experienced:Chinese)	Coef β	SE(β)	**z**	**p**
(Intercept)	0.69	0.17	4.1	**<.001**
imitated	−3.72	0.10	−39.1	**<.001**
Korean	−0.09	0.07	−1.3	>0.2
Macedonian	−2.10	0.08	−25.9	**<.001**
Turkish	−3.91	0.13	−29.6	**<.001**
inexperienced	−1.22	0.18	−6.7	**<.001**
imitated:inexperienced	2.58	0.11	24.2	**<.001**
Korean:inexperienced	0.59	0.09	6.8	**<.001**
Macedonian:inexperienced	1.26	0.10	12.7	**<.001**
Turkish:inexperienced	2.37	0.15	15.9	**<.001**

Table 2.2: **Fixed effects with coefficients and *p*-values for 'yes' responses by accent and experience.** Reference levels: accent: Chinese, experience: Experienced

voice. These listeners also appear to be more likely to respond 'yes' to an authentic Chinese accent than are inexperienced listeners. Inexperienced listeners, by contrast, appear to be more likely than experienced listeners to identify an imitated Chinese accent as 'authentic'. Given the large number of data points in this experiment, it is highly likely that even small, uninformative differences will achieve statistical significance. However, the magnitude of these two differences, and their usefulness in clustering and classification below, suggest that these differences are not merely significant but also meaningful. This pattern of responses suggests that experienced and inexperienced listeners are employing different strategies when deciding whether a particular voice is 'authentic'.

Subject and Item were included as random effects in a generalized linear mixed model with binomial errors and a logit link function. The dependent measure in this model is whether the participant responded 'yes' to the stimulus ('yes' response regardless of accuracy is an indicator of the listener's belief that the stimulus is authentic Chinese). Accent and Experience and the interaction of Accent with Experience were included as fixed effects in the model. Factor levels, coefficients, standard error, *z*-score, and *p*-values for each level of these factors and interaction are reported in table

(ref. level: experienced:imitated)	Coef β	SE(β)	**z**	**p**
(Intercept)	−3.03	0.18	−16.7	**<.001**
Chinese	3.72	0.10	39.1	**<.001**
Korean	3.63	0.10	35.1	**<.001**
Macedonian	1.62	0.11	15.2	**<.001**
Turkish	−0.19	0.14	−1.3	>0.2
inexperienced	1.36	0.20	6.9	**<.001**
Chinese:inexperienced	−2.58	0.11	−24.2	**<.001**
Korean:inexperienced	−1.99	0.12	−16.9	**<.001**
Macedonian:inexperienced	−1.32	0.12	−10.7	**<.001**
Turkish:inexperienced	−0.21	0.16	−1.3	>0.2

Table 2.3: **Fixed effects with coefficients and *p*-values for 'yes' responses by accent and experience.** Reference levels: accent: imitated, experience: Experienced

2.2.

Chinese was the default reference level for the Accent factor and experienced was the default reference level for the Experience Factor in the reported comparisons. We therefore interpret the reported results for the 'imitated' factor level of the Accent variable with respect to responses to authentic Chinese-accented stimuli by experienced listeners. In other words, this row is a test of the significance of the apparent trend, visible in figure 2.3, for experienced listeners to respond 'yes' more to authentic Chinese more reliably than to imitated Chinese; this difference is significant ($\beta = 3.72, p < 0.001$). Experienced listeners were significantly more likely to respond 'yes' to authentic Chinese stimuli than to imitated Chinese (or to any other accent). Experienced listeners were also significantly more likely to respond 'yes' to authentic Chinese-accented stimuli than were their inexperienced counterparts ($\beta = 1.22, p < 0.001$).

Table 2.3 reports output for the same model but rotated so that 'imitated' is the default reference level for the Accent variable. We now interpret the 'inexperienced' level of the Experienced variable with respect to responses by experienced listeners to imitated stimuli. Inexperienced listeners are significantly more likely than experienced listeners to respond ‘yes’ to imitated Chinese stimuli ($\beta = 1.36, p < 0.001$).

Re-running the model with the default level of the Accent variable switched to Korean reveals that experienced listeners were statistically more likely to respond ‘yes’ to an authentically Korean-accented voice ($\beta = -0.62, p < 0.0008$). It may well be that Experienced listeners are responding to a percept of ‘authenticity’ in the Korean voice in their desire not to make any false negative responses –though this is highly speculative.

The apparent difference between responses to Macedonian-accented English is not significant ($\beta = 0.04, p = 0.83$), while the Turkish accented-English difference is significant ($\beta = 1.1443, p < 0.001$).

2.4.5.2 Accuracy

Figure 2.4 is quite a different view of the data than the proportion ‘yes’ visualization in figure 2.3. Here, rather than the height of each bar being proportional to other bars in an experience level, bar height is simply the number of correct responses over the number of total responses to an Accent for each experience level. Here we can see that, as anticipated, experienced listeners appear to be more accurate when responding to either authentic Chinese or imitated Chinese stimuli. However, accuracy results do not, on their own, necessarily reveal the listeners’ ability to detect a signal such as the Chinese accent in this experiment. A listener hoping to have perfect recall on the Chinese-identification task could, for example, simply press the ‘yes’ button in response to each stimulus item. Overall performance would be poor, but accuracy to the Chinese stimuli would be perfect.

A measure of response sensitivity from signal detection theory, d’, represents the distance between a listeners’ ability to maximize hit rate (correct identifications) and minimize false rejections. Table 2.4 reports Hit and False Alarm rates in these results

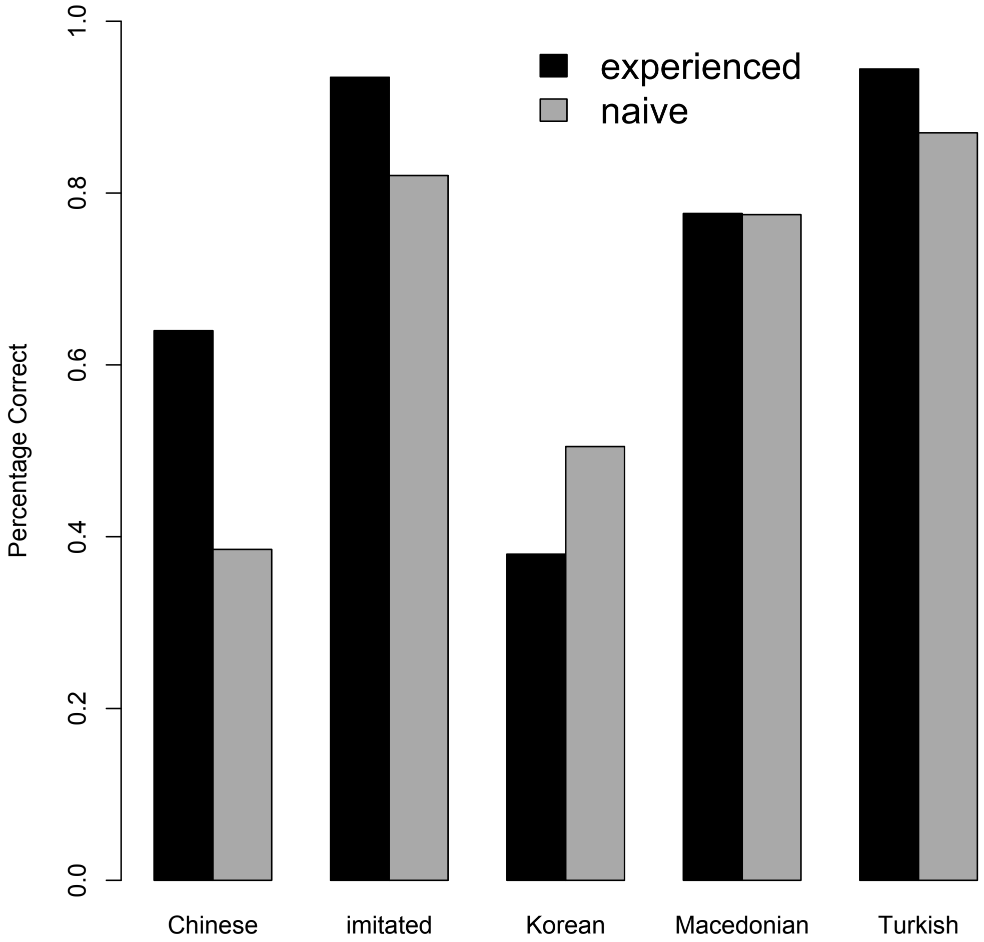

Figure 2.4: **Correct responses by accent and experience.** Accuracy bars are not proportional across experience level or within language.

along with d' and criterion (c) scores. The question addressed by these metrics is the extent to which listeners are correctly identifying authentic Chinese and rejecting other accents. The criterion measure, or c, is a measure of response bias that attempts to model the decision criterion chosen by listeners when completing a task.

$$d' = z(H) - z(F) \tag{2.1}$$

Experience	Hit Rate	False Alarm Rate	d'	c
experienced	0.6398055	0.2059191	1.1786019	0.2313620
inexperienced	0.3852056	0.2418758	0.4084442	0.496052

Table 2.4: **Signal detection results for Experiment 1**

$$c = -0.5[z(H) + z(F)) \tag{2.2}$$

H represents the hit rate: correct 'yes' responses divided by possible 'yes' responses (equivalent to recall in information retrieval). F represents the false alarm rate: incorrect 'no' responses divided by the number of potentially correct 'no' responses). z() is a z-transform function (taking probabilities and returning z-scores). Positive c scores correspond to a tendency to respond 'no' during the task. Both groups of listeners are biased to respond 'no' but experienced listeners much more weakly so. If c = 0 the listener is unbiased; naive listeners have a stronger 'no' bias c = .496052 than experienced listeners c = .2313620.

2.4.5.3 Clustering

Figures 2.5 and 2.6 show a visualization of a correspondence analysis of the yes/no task data. Correspondence analysis is an unsupervised clustering technique. From a contingency table of 'yes' responses by participants to each level of the language factor, two separate square distance matrices are calculated: a row x row matrix (in this case, distances between participants) and a column x column matrix (distances between languages). The software used here, Baayen's `LanguageR` package for the `R` open source statistical environment (R Development Core Team, 2011), uses a chi-squared distance metric. Like principal components analysis does for real-valued data, correspondence analysis provides a low-dimensionality map of both rows and columns in a contingency table (Baayen, 2008). 'Factor 1' on the x-axis represents

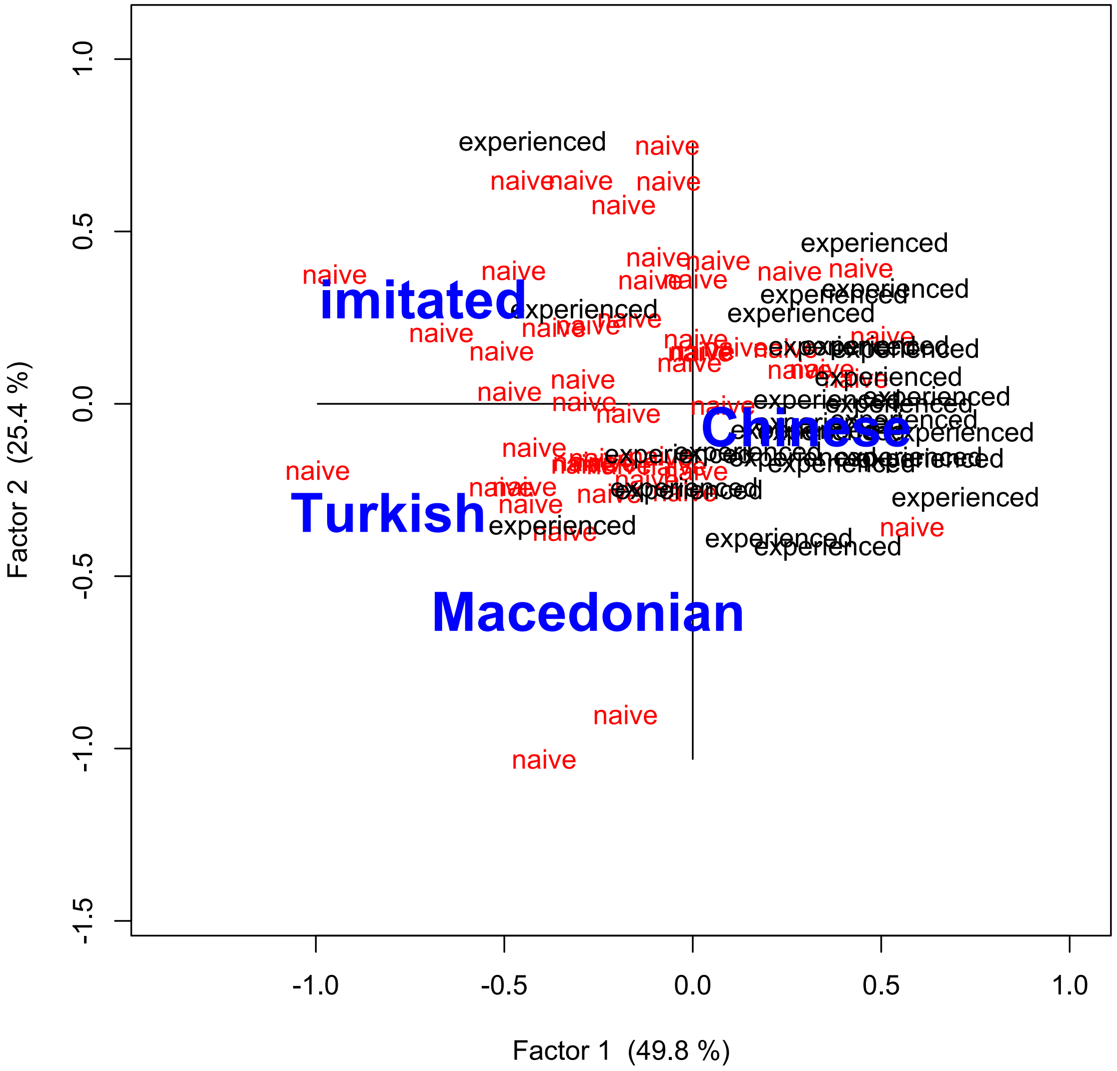

Figure 2.5: **Correspondence analysis of experienced and inexperienced listeners** Experienced listeners cluster tightly around the authentic Chinese target while naive (inexperienced) listener responses are more diffuse.

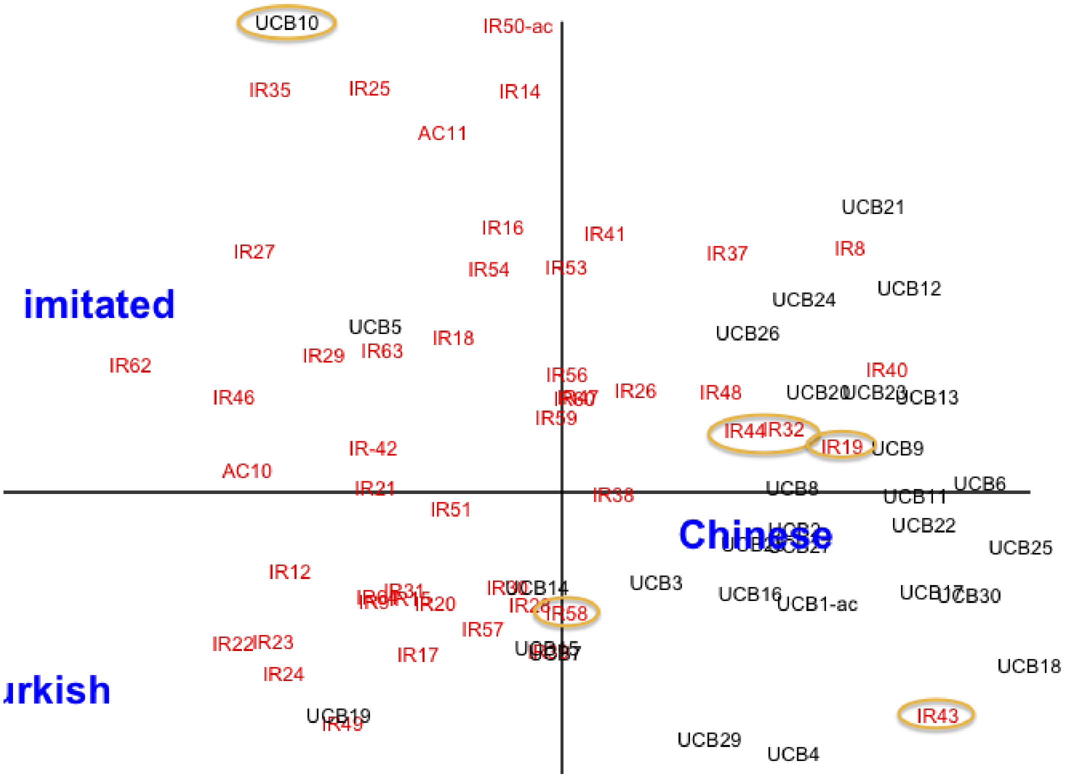

Figure 2.6: **Correspondence analysis of experienced and inexperienced listeners cropped and zoomed to highlight participant ID detail.** Circled participant IDs were independently excluded prior to further data analysis; 'ucb' indicates experienced listeners and all others are from the inexperienced group.

the most informative column, authentic Chinese, with an eigenvalue rate of 0.4984 or 49.84% of the variance in the table. ‘Factor 2’ on the y-axis represents the second most informative column, imitated Chinese, with an eigenvalue rate of 0.2538 or 25.38% of the variance in the table. This two-dimensional visualization of the data captures roughly 75.2% of the variance in the table; Korean contributed virtually no explanatory power to the map and has dropped out of the visualization.

Intuitively from figure 2.5 we can see that the experienced Heritage Mandarin participants from the University of California, Berkeley have, for the most part, clustered tightly around the Chinese label. This suggests that, as predicted, these listeners were more attracted to Chinese for responses of ‘authentic Chinese’ than to any other language. The clustering of inexperienced (here rendered as ‘naive’ for visual differentiation) monolingual English participants from the University of Michigan is much more diffuse. They appear to be attracted to both the imitated and authentic Chinese languages for ‘yes’ responses with neither cluster being a particularly good predictor of inexperience.

Interestingly, all but one of the participants who were independently excluded from data analysis are outliers on this plot. Figure 2.6 is a zoomed and cropped view with the excluded participants circled. Participant UCB10 was excluded from the experienced data set for misrepresenting his identity and, reassuringly, is among the most inexperienced of the inexperienced participants in terms of attraction to the imitated Chinese voices. Participants IR19, IR32, IR43, IR44 and IR58 were excluded from the inexperienced data set for self-reporting extensive or Heritage experience with Mandarin-accented English. Of these, only IR58 does not clearly cluster with the experienced participants.

		Prediction
Label	experienced	inexperienced
experienced	22	8
inexperienced	6	47

Table 2.5: **Cross tabulation of iterative SVM classification results.** Training set included all data with a single participant withheld as test data at each iteration. Light gray cells indicate mismatches between predicted classification and data label.

		Prediction
Label	experienced	inexperienced
experienced	21	7
inexperienced	2	45

Table 2.6: **Cross tabulation of iterative SVM classification results.** Training set included only non-excluded data with a single participant withheld as test data at each iteration. Light gray cells indicate mismatches between predicted classification and data label.

2.4.5.4 Classification

The visualization in figure 2.5 is enlightening about the structure of the data and suggests that this task may have some predictive power. But the data's usefulness for classification purposes is less clear. Taking the two principal components identified in the correspondence analysis, a support vector machine (SVM) classifier was trained with the full data set. A support vector machine uses a supervised learning algorithm to find a hyperplane that divides elements of a labelled training data set as cleanly as possible into, for a binary classification problem, two groups that are as distinct as possible. To test the classification of each participant, a separate classifier was trained on the entire data set with only that participant withheld as test data. The classification of that withheld test participant was then predicted using this model.

This was repeated for all participants using the built-in **svm()** command in **R**.

The contingency table presented in table 2.5 cross-tabulates SVM-predicted labels against labels from the data. Values in the light gray cells represent mismatches between the SVM-predicted classification of the test item and the item's original category label in the data. The SVM classifier correctly predicted the original labels in table 2.5 in 73.3% of cases for experienced listeners and in 88.7% of attempts to classify inexperienced listeners. I hesitate to refer to these as classification errors because the initial label assignment in the data, participants' self-reported experience levels and the predictions of the classifier, each introduce the potential for error. In the absence of more information, it is difficult to determine which label more closely resembles objective reality.

We have independent evidence that several of the participants included in the testing for table 2.5 were incorrectly labelled in the original data. Table 2.6 presents the results of training and testing the SVM classifier with these suspect data withheld. Given these slightly more accurate data, the SVM classifier correctly predicted the label for experienced listeners in 75% of the tests. The label for inexperienced listeners was correctly predicted in 95.7% cases. It is not possible to know if the remaining classification mismatches are due to mislabelings in the source data or to true classification errors.

2.4.5.5 Reaction Time

For the sake of completeness I report the reaction time data. I predicted that experienced listeners should have lower reaction time latencies overall. The question, after all, is whether the voice is authentic Chinese and these listeners have experience to draw on. This prediction was not upheld.

As is generally the case, the reaction times were logarithmically distributed and a log transform was needed to more closely approximate the normal distribution.

Reaction times longer than two standard deviations above the mean (10436.85ms) were excluded from analysis.

The data were analyzed using a linear mixed effects model in which logRT was the dependent variable. Accent, Experience, and the interaction of Accent x Experience were included as fixed effects in the model. Participant and Item were included as random effects. R's linear mixed effects utility, `lmer()` from the `lme4` package (Bates et al., 2011), does not calculate *p*-values directly when not using the binomial link function. Instead, standard practice is to use Markov Chain Monte Carlo simulation to estimate *p*-values. The values reported in table 2.7 were estimated using the `pvals.fnc()` command in the `languageR` package (Baayen, 2008).

The log transformed reaction time data are plotted by accent and experience level in figure 2.7. Inexperienced listeners were, on average, faster than experienced listeners when responding. This was particularly true when the stimulus accent was authentic Chinese, when inexperienced participants' response latencies were, on average, 455ms faster than those of experienced participants. This difference is significant ($\beta = -0.13, p < .001$).

(ref. level: experienced:Chinese)	Estimated β	SE(β)	**z**	**p**
(Intercept)	8.06	0.071	113.55	**<.001**
inexperienced	-0.13	0.018	-6.84	**<.001**
imitated	0.19	0.007	28.85	**<.001**
Korean	0.25	0.008	29.28	**<.001**
Macedonian	0.017	0.008	1.98	**0.0473**
Turkish	-0.06	0.008	-7.77	**<.001**
inexperienced:imitated	0.06	0.008	7.59	**<.001**
inexperienced:Korean	0.01	0.010	1.16	0.2458
inexperienced:Macedonian	0.03	0.010	3.30	**0.0010**
inexperienced:Turkish	0.04	0.010	3.50	**0.0005**

Table 2.7: **Fixed effects with coefficients and estimated *p*-values for log transformed reaction times by accent and experience**

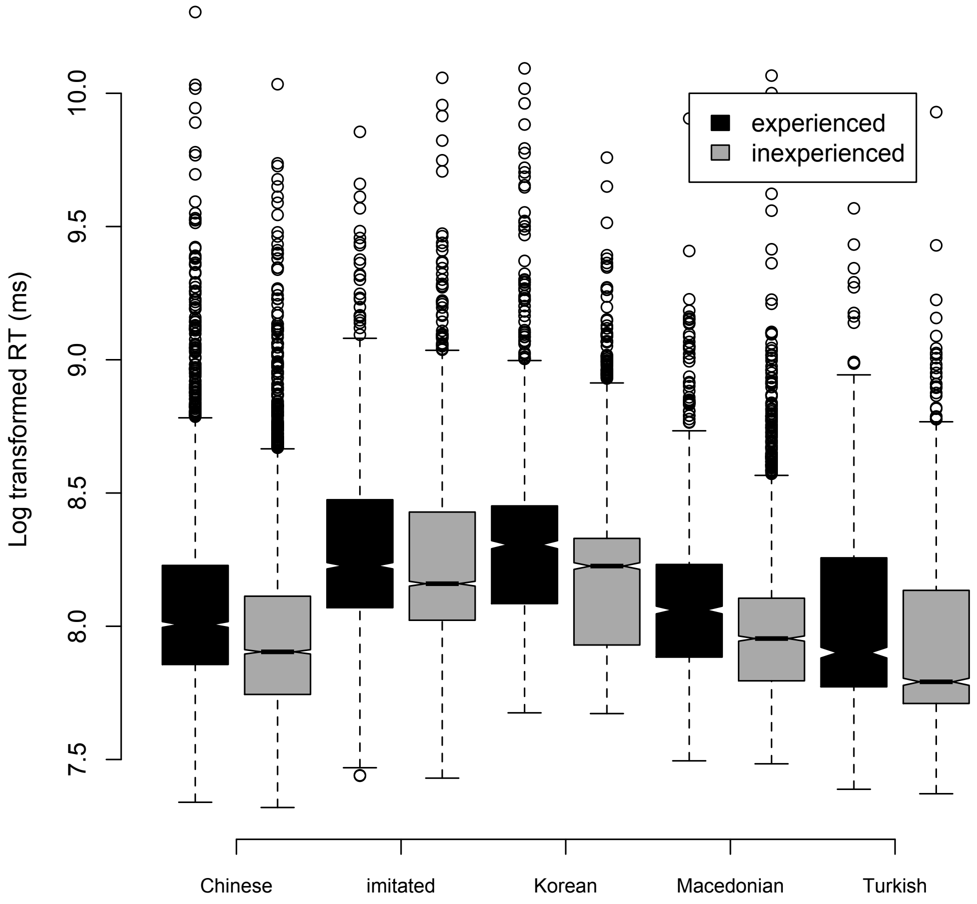

Figure 2.7: **Log transformed reaction times by accent and experience.**

2.5 Discussion

If speech perception research has established anything with absolute certainty it is that speech is a complex and variable phenomenon. Identity is no less complex and no less variable. The study of listeners' use of socioindexical knowledge during speech perception therefore takes on the daunting challenge of studying the interaction of these topics. My purpose in this chapter has been to lay out some of the sources of variability speakers may be aware of, to take seriously the scholarship fields outside speech perception have to offer on the topic of identity and to explore ways of reducing the dimensionality of the identity problem for the exigencies of laboratory work.

The primary goal of the experiment presented here was to quantify the extent to which listeners with little or no experience listening to native speakers of a target variety nevertheless use socioindexical knowledge in a systematic way during perception and to compare their performance to that of experienced listeners. Inexperienced and experienced listeners, at least of Mandarin-accented English, have differentiable behavioral responses to authentic Chinese and imitated Chinese stimuli.

It seems reasonable to infer that this means listeners are drawing on different forms of knowledge when detecting an authentic Chinese accent. This finding does not refute exemplar models, per se. It may not even be especially problematic for exemplar theoretic models of speech perception. The experienced listeners could well be drawing on stored episodic traces of experience with authentic Chinese-accents while inexperienced listeners draw on stored episodic traces of comedians and actors imitating the accent, generalizations from other Asian languages, etc. Nor is it the goal of this dissertation to refute exemplar models. It would suggest, however, that interpreting behavioral results like those in Staum Casasanto (2009a) as clear evidence that listeners have stored experiential knowledge is premature and very likely underestimates the complexity and nuance of listeners' use of socioindexical expectations during speech perception.

Inexperienced listeners are less accurate than experienced listeners at identifying an authentic Chinese voice, but their performance is not zero or chance. They are successfully drawing on *some* kind of knowledge and this knowledge may well be the same stereotypes they draw on when, incorrectly, identifying imitated Chinese, imitated French, or imitated American English as 'authentic'.

It seems to me that it is quite reasonable to imagine that Staum Casasanto (2009a)'s listeners, for example, were drawing on stereotypical knowledge of African American English rather than rich stores of experience listening to speakers of AAE when they completed the mass/mast sentence completion task. This makes her result no less fascinating and no less compelling, but it does suggest that the situation is even more interesting. If listeners can systematically draw on either first-hand experience with a variety or on higher level ideologies regarding that variety to reduce the complexity of the speech perception task, they surely will. In the following chapter I investigate the extent to which both inexperienced and experienced listeners can benefit from socioindexical information about a purported speaker when understanding and transcribing speech-in-noise.

An additional goal was to explore the usefulness of this task as a means of directly estimating participants' experience level with authentic Chinese-accented English. Whereas the IAT test offers a means of understanding listeners' unstated, and possibly unconscious, biases, the experiment presented here attempts to measure listeners' experience with a particular language variety. In this goal I believe I have been somewhat successful. The *yes/no* task presented here is extremely easy to build and administer for any target variety, requiring only a fairly small set of native and imitated language recordings and access to a participant population. I believe this task will be particularly helpful in gauging listeners' experience with language varieties for which the inexperienced and experienced populations are not so easily identified as Mandarin-accented English. It could also be helpful assessing listeners' experience with varieties that they may have ideological reasons to disavow knowledge of (e.g. middle class African American students might wish to distance themselves from knowledge of AAE –particularly in a formal context).

This last example raises an important caution about the interpretation of the present result. Though every attempt was made to keep the experiment as consistent as possible despite the change of venues, the inexperienced and experienced listeners are fundamentally not performing the same task. Though they used the same computers, were given the same instructions, heard the same stimuli, saw the same sentences and performed the same physical task, experienced listeners were inescapably aware of having been recruited precisely because they were Heritage speakers of Mandarin Chinese. The inexperienced listeners were simply asked to identify the authenticity of a non-native accent. Experienced listeners, by virtue of their identity, are not only trying to identify the authenticity of an accent but, in a very real way, are also striving to demonstrate their own authenticity. I believe this difference alone accounts for experienced listeners' significantly slower reaction times (section 2.4.5.5) on the authentic Chinese items. Future uses of this technique will need to be more careful about keeping experienced participants naive to the role of their experience in the experiment.

Finally, this experiment lays the groundwork for a larger project exploring the acoustic correlates of 'Chinese-ness' for both experienced and inexperienced listeners. The next step for this project is to attempt to identify those features. This is equally necessary for experienced listeners' responses as it is for inexperienced listeners.

CHAPTER III

Socioindexical Expectations and Speech Perception in Noise

Chapter II investigated the differential abilities of experienced and inexperienced listeners to distinguish between authentic Chinese, imitated Chinese, and other non-native English accents. It is clear that experienced listeners are better able to identify an authentic Chinese accent and less likely to be fooled by imitated Chinese. However, inexperienced listeners –listeners with no reported experience with authentic Chinese accents– perform at better than chance levels on the identification task. Inexperienced listeners are clearly able to draw on some set of expectations when performing the accent identification task.

The present chapter turns to an investigation of the much more fundamental question of whether listeners can use the expectation of an accent in a systematic way to aid in speech perception and word recognition.

3.1 Experiment 2: The Perception of Non-Native Speech in Noise

As briefly described in Chapter I, there is a growing body of work, much of it inspired by exemplar theories, investigating listeners' use of social knowledge during speech perception. One criticism to be levied against much of this existing work on the influence of social knowledge during speech perception is the use of unnatural, off-line, or highly meta-linguistic experimental tasks.

It seems clear that these tasks provide interesting information about the influence of social cognition on language processing at some level. What is less clear from these indirect evaluations is whether the influence is actually on, as is generally claimed, the low level processing of fine phonetic detail.

Niedzielski (1999), for example, conducted an influential study of the influence of social information on the perception of sociolinguistically informative vowels. Niedzielsi played a series of sentences for a group of listeners from Detroit, Michigan who were asked to concentrate on the vowel in a particular word in each sentence. These listeners then selected, for each sentence, one of six isolated resynthesized[1] vowels (see also Willis, 1972) as the best match for the target word.

As a general rule, Detroit speakers can be expected to participate in the Northern Cities Chain Shift and Canadian Raising (Labov, 1994). Prior to participating in the task listeners were told either that the speaker of the target sentences was Canadian or that she was, like themselves, from Detroit. For the Canadian Raising vowels, Detroit listeners were much more likely to choose a synthesized sample that is consistent with the acoustics of the target word when they believe the speaker to be from Canada. Listeners who believed the speaker to be from Detroit instead chose resynthesized vowels consistent with unraised or more "standard" vowels. Results across groups were much more consistent for the Northern Cities Chain Shift targets with perhaps a slight trend, not statistically significant, toward unshifted variants when the listeners believe the speaker to be from Detroit.

Niedzielski interprets her results as evidence of the influence of social information on the processing of fine phonetic detail. Being asked to focus on a particular vowel in

[1] Due to the nature of resynthesis, the participants in this task, when responding to stimuli with diphthongs, were hearing only non-varying monophthongal resynthesized "diphthong onsets".

a particular word is itself a peculiar, meta-linguistic task. Then being asked to select a similar vowel from a set of six examples separates the listener still further from the moment of perception. This task does not directly test listeners' ability to process fine phonetic detail but, instead, tests how listener beliefs about speaker identity can influence the way they choose a synthetic vowel from a list. We simply don't know how accurately or inaccurately listeners were perceiving the phonetic details of the target words. What we really know is which synthesized targets sound 'Canadian' and which sound 'Detroiter'. We also know that, apparently, Detroiters have an accurate stereotype of Canadian participation in Canadian Raising but, at least in 1999, were not yet aware that this change was also underway in their own speech and the speech of their fellow Detroiters. This work is an interesting and useful investigation of the influence of language ideologies on the identification of steady-state vowel targets along the lines of Cynthia Clopper's work on dialect perception (e.g. Clopper, 2004).

Hay et al. (2006a) and Hay and Drager (2010), described in more detail in Chapter I, use essentially the same task, with Australian and New Zealand vowel targets, to make the still stronger claim that this behavioral result should be interpreted as evidence for exemplar theories of speech perception. It is unclear how strongly this interpretation is motivated by the data or the task. It seems clear only that listener ideologies about sociolinguistic stereotypes can influence performance on a vowel selection task. The task would need to be much more natural and also have fewer intervening cognitive steps between the presentation of the acoustic stimuli and the behavioral response being measured to motivate claims about the processing of fine phonetic detail.

Evidence for the influence of social knowledge on speech perception would be more compelling if the tasks involved either were closer to listeners' daily experience with language, more closely restricted measures of success to the processing of phonetic detail or word recognition, or both. One can imagine, for example, attempting to replicate any of these three studies using an AXB task in which all listeners hear precisely the same recordings and are asked to decide whether the second word in each triplet sounds more like the A target or the B target. There is not a good scale for ranking the naturalness of a given task, but deciding which of two alternatives an entire word is more like seems, while not especially 'natural', at least intuitively more like normal language use than choosing a vowel from a list of six alternatives.

The present task is intended to be more natural than the tasks described so far. In what is essentially a socioindexical priming experiment (c.f. Bruce, 1958), listeners are asked to listen carefully to a series of audio recordings and to transcribe what they hear in standard English orthography. This task seems intuitively quite natural and linguistic. Transcription, in the form of note taking, is not outside the realm of undergraduate experience, and the noise chosen to raise the difficulty of the task (described in detail in 3.2) was multitalker babble to enhance the ecological plausibility of the task.

The task is still 'off-line' in the sense that most listeners do not begin typing until they have finished hearing the entire sentence. However, the simplicity of the task does seem to restrict the sources of error to semantic reinforcement (which has been controlled for), listener idiosyncrasies, or the processing of fine phonetic detail and word recognition. There is no familiarization or training stage required for this particular experiment (although there are practice items), the instructions were brief and carefully scripted, and interpretation of the results seems relatively straightforward.

In spirit, this experiment is intended to be a more natural, more real-time extension of Rubin (1992), which was discussed at length in Chapter I. In that experiment, participants heard recordings of Standard American English speech and images were used to manipulate their beliefs about the racial identity (and thus native language) of the speaker. Similarly, in this experiment, listeners hear recordings of Chinese-accented English and different faces are displayed to shift socioindexical expectations during transcription. The alignment of matching/mismatching face and voice pairs has been inverted from Rubin's design. Specifically, in Rubin (1992), those seeing an Asian face expected an accent that was not present in the audio recordings. Those seeing a Caucasian face did not expect a foreign accent and did not hear one. In the present study, those seeing an Asian face expect an accent and the voice in the recorded sentences does, indeed, have a Chinese accent. Listeners seeing a Caucasian face will not anticipate an accent but will, nevertheless, hear one.

If listeners in the present experiment who believe the speaker to be Chinese transcribe identical recordings more accurately than those who believe the speaker to be Caucasian then it seems fairly clear that, contrary to Rubin and Lippi-Green's interpretations of Rubin (1992), expectation of a foreign accent can have a facilitatory effect on the understanding of accented speech. What implications this effect may or may not have for the usefulness of social *knowledge* during speech perception will be the central question of the subsequent discussion. A silhouette condition, intended to convey no socioindexical information, is included to help distinguish between facilitation when the face and voice support one another and inhibition when there is a face/voice mismatch –a control missing from the Rubin study.

3.2 Methods

3.2.1 Stimuli

Stimulus materials consisted of 30 pairs of high and low predictability sentences originally developed by Bradlow and Alexander (2007) for presentation to non-native English speakers. Bradlow and Alexander created the high predictability sentences using an iterative sentence completion paradigm with groups of non-native and native speakers of English. Sentences in the high predictability list are those that consistently received the most consistent completion results from both populations. The

High Predictability	**Low Predictability**
Elephants are big animals.	He pointed at the animals.
A pigeon is a kind of bird.	We pointed at the bird.
The war plane dropped a bomb.	Dad talked about the bomb.
A quarter is worth twenty-five cents.	He pointed at the cents.
We heard the ticking of the clock.	She looked at the clock.
The team was trained by their coach.	We read about the coach.
Many people like to start the day with a cup of coffee.	Mom pointed at the coffee.
February has twenty-eight days.	There are many days.
Last night, they had beef for dinner.	He talked about the dinner.
My parents, sister and I are a family.	We read about the family.
A race car can go very fast.	She thinks that it is fast.
The good boy is helping his mother and father.	Mom pointed at his father.
People wear shoes on their feet.	Mom looked at her feet.
When sheep graze in a field, they eat grass.	Dad pointed at the grass.
I wear my hat on my head.	She pointed at her head.
At breakfast he drank some orange juice.	Mom looked at the juice.
In spring, the plants are full of green leaves.	She talked about the leaves.
People wear scarves around their necks.	She talked about their necks.
For dessert, he had apple pie.	Mom talked about the pie.
She made the bed with clean sheets.	Dad talked about the sheets.
Rain falls from clouds in the sky.	Dad read about the sky.
The sport shirt has short sleeves.	He looked at the sleeves.
Football is a dangerous sport.	This is her favorite sport.
A book tells a story.	We looked at the story.
A wristwatch is used to tell the time.	This is her favorite time.
Birds build their nests in trees.	He read about the trees.
He washed his hands with soap and water.	We talked about the water.
Monday is the first day of the week.	This is her favorite week.
Bob wore a watch on his wrist.	He looked at her wrist.
The color of a lemon is yellow.	Mom thinks that it is yellow.

Table 3.1: **High/Low predictability sentence pairs from Bradlow and Alexander (2007)**

low predictability sentences replace the semantically informative material with uninformative frames. These sentences were selected for the transcription task for three reasons. First, the keywords have been normed by Bradlow and Alexander for recognizability by non-native speakers. Second, the pairing of high and low predictability sentences should allow us to gauge any contribution of social knowledge to sentence perceptibility over and above the better-understood contribution of semantic knowledge. Third, the Wildcat Corpus (Van Engen et al., 2010) contains high quality recordings of these sentences by a number of native Mandarin speakers. The recordings used in this experiment were read by a 23 year old female Chinese native speaker of Mandarin (Wildcat Corpus speaker CHF02). The sentences are listed in table 3.1.

The scripted recordings from the Wildcat Corpus were segmented into individual sentence-length files and equated in amplitude. These files were then mixed with native English multi-talker babble (Van Engen and Bradlow, 2007) using the sox audio processing tool to create speech-in-noise recordings with a +4 dB signal-to-noise ratio at the target word. This signal-to-noise ratio was determined after a series of pilots using the full set of sentences with no noise in which participants across conditions demonstrated a clear ceiling effect in transcription accuracy. An informal listening task completed by several researchers unfamiliar with the semantic content of the sentences suggested that mixing 76 dB noise with a 72 dB signal would result in sufficient transcription errors for the purposes of the experiment.

Multi-talker babble was selected over white, Brownian, or other possible types of noise to enhance the ecological plausibility of the stimuli for participants. Listeners in this task are being asked to draw on their socioindexical expectations under laboratory conditions; these more random types of noise created stimuli that seemed, to the experimenter, to be more clinical and less natural-sounding.

The target words themselves occur uniformly in sentence-final position with the falling intonation typical of English declaratives and with the declination typical of the end of a prosodic group. This speaker was chosen from the set of available speakers, in part, because there is no obvious list intonation in her reading of the scripted sentences. Beyond this uniformity the target items represent a rather varied set of vowels, consonants, consonant clusters, number of syllables, and morphological complexity.

It is worth noting that the actual target norms for L2 English speakers in China have traditionally been British rather than American English (Kirkpatrick and Zhichang, 2002). Though there may be a shift underway currently to American English norms in textbooks and pedagogical recordings, these materials have traditionally featured British English (Xinting Zhang, personal communication, June 17, 2011). This fact surely influences the English acquired by Chinese learners and may interact with and shape American listener expectations about Chinese-accented English. The belief that a speaker of Chinese English will be non-rhotic, for example, may well be attributable to this legacy.

Prior to the presentation of the experimental stimuli, listeners heard and transcribed four practice items intended to capitalize on recognizable associations between face, accent/voice, and semantic content. The goal of these practice items was both to make participants comfortable with the transcription user interface and, implicitly, to reinforce the illusion that face and voice would be somehow meaningfully linked in the experiment. Listeners transcribed, in random order, two recordings of Leonard Nimoy as the character Spock and two recordings of Arnold Schwarzenegger speaking characteristic lines of dialogue presented in multitalker babble.

3.2.2 Visual Stimuli

Like most of the experiments discussed in the introduction, the present experiment is an inverted matched guise task. Matched guise is a well-established experimental technique in sociolinguistics for teasing apart auditory indexical information (Lambert, 1960) and perceived socioindexical properties. The present experiment is ‘inverted’ because it presents visual stimuli to establish socioindexical expectations and then measures the extent to which these socioindexical expectations can influence the perception of phonetic detail and word recognition in noise.

One of three images was presented to listeners to establish these socioindexical

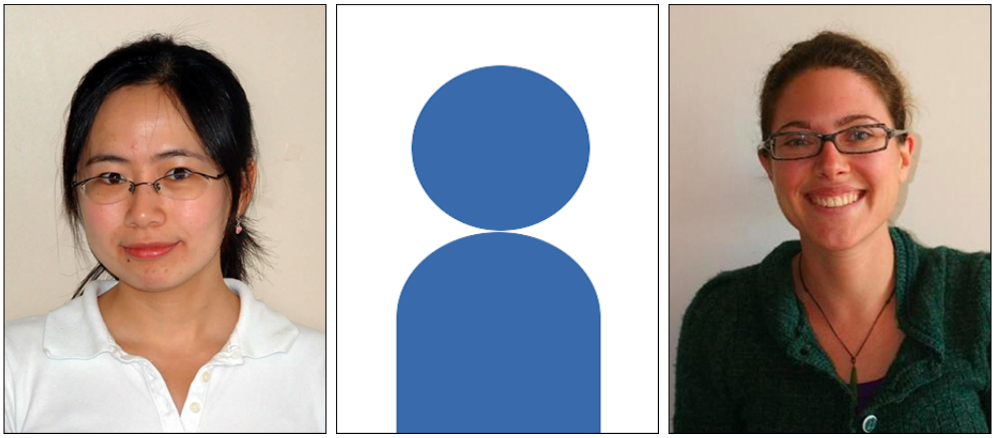

Figure 3.1: **Faces used in the transcription experiment**

expectations; these faces are shown in figure 3.1. Each listener saw only one of the three images (between-subjects design) and the image was displayed for the duration of the trial. The Asian and Caucasian images were found via web search for license-free portraits and, beyond an informal survey of several graduate students in Linguistics, have not been formally normed for attractiveness, racial typicality, gender stereotypicality, memorability, etc. at the time of writing.

3.2.3 Participants

As in Experiment 1, eighty-seven undergraduate students participated at one of two experiment sites: the University of Michigan phonetics lab or the University of California, Berkeley phonology lab.

3.2.4 Inexperienced Listeners

The 57 inexperienced listeners reported in Experiment 1 completed the present task prior to their participation in the authentic Chinese accent identification task and completion of a brief language history survey. The same 7 participants excluded in Experiment 1 were also excluded for this experiment. Data for 50 participants are reported here: 16 in the Asian face condition, 16 in the Caucasian face condition, and

18 in the silhouette condition.

3.2.5 Experienced Listeners

The 30 Heritage Mandarin-speaking listeners reported in Experiment 1 also completed the present task prior to their participation in that task. In addition to the two discarded subjects described above, a third data file was overwritten prior to analysis due to experimenter error. In all, 10 participants were randomly assigned to the Asian face condition (one missing), 8 to the silhouette condition, and 10 to the Caucasian face condition. Time constraints limited the number of participants who could be engaged in the study.

3.2.6 Procedure

Inexperienced listeners used Apple Macbook Computers (model 4,1; late 2008) in an IAC sound-attenuated booth at the University of Michigan, Department of Linguistics; stimuli were presented over AKG K271 mkII headphones.

Experienced listeners participated in the University of California, Berkeley's phonology laboratory rather than at the University of Michigan. This is a quiet space dedicated to speech perception experiments, but is not a sound-attenuated booth. AKG k240 headphones were used in place of the AKG K271 mkII.

Prior to their arrival, participants were randomly assigned to one of the three Face conditions: Asian face, Silhouette, or Caucasian face. Responses were entered via the Macbook keyboard and listeners were instructed to advance trials using the return key to minimize trackpad use. Stimuli were presented using Superlab stimulus presentation software version 4.0.8. Volume was set at a comfortable listening level.

The exact instructions provided were:

> This experiment is designed to help us understand what information listeners like yourself use when transcribing speech in noise. During the

> experiment you will hear 60 sentences. Your task is simply to type, as carefully as you can, what you hear. You can only listen to each sentence once –they can not be replayed– so please listen closely. If you are unable to make out all of the words in the sentence please type the words you *are* able to understand. Your task is made somewhat harder than it sounds by the presence of what is called multitalker babble, you may also have heard the term 'cocktail party noise'. The words you are listening for are embedded in the sound of many other people speaking at the same time. There will be four practice sentences for you to hear what the noise sounds like and to get comfortable using the program. Please take your time; there is no rush. Spelling does not count, but please try to type carefully. Simply press return when you have finished typing to advance to the next sentence. Do you have any questions?

3.2.7 Predictions

If exemplar theories of speech perception (e.g. Johnson, 2006) are correct about the role of social knowledge in the processing of fine phonetic detail, and if the prevailing interpretation of recent findings in sociophonetic perception is correct, then our predictions are clear. We should see a shift in listeners' responses suggesting that listeners in the different socioindexical conditions are processing incoming acoustic information using a different set of subcategorical, phonemic and lexical expectations. This is a change equivalent to replacing the acoustic model in an automatic speech recognition system. Listeners who have some knowledge or experience with Chinese-accented English will have the base activations of their Chinese-accented exemplars raised. Another way of stating this prediction, without the assumption of stored episodic traces of previous linguistic experience, would be that the listeners' prior probabilities over subcategorical, phonological and lexical forms will shift to favor the retrieval of those forms consistent with Chinese-accented English. Listeners, like those selected for participation in this experiment, with little or no experience with Chinese-accented English should perform identically on the transcription task regardless of face.

However, since even the most inexperienced listeners in the yes/no task described in Chapter II were capable of better-than-chance performance identifying an authentic Chinese accent, we have reason to suspect that this strong prediction will not be upheld. Inexperienced listeners are apparently drawing on some kind of knowledge of Chinese –either stereotypical knowledge of the accent or ambient cultural exposure is greater than listeners estimate. Therefore, I predict that even inexperienced listeners will see some facilitatory effect of the Asian face. Experienced listeners should be both more accurate transcribers of Chinese-accented English overall and, with their greater experience, should show a larger benefit of socioindexical knowledge than the inexperienced listeners.

Across both groups of participants, though, transcription should be most accurate given the Asian face, least accurate given the Caucasian face (potentially due to mismatch-induced inhibition), and the silhouette, with no socioindexical information, should hover between the two conditions.

Semantic knowledge is a powerful tool for disambiguating speech in noise and, will, I predict, overwhelm any facilitatory effect of face. Facilitation should therefore be strongest in the Low predictability sentences where the information provided by socioindexical knowledge can provide the most benefit.

3.2.8 Results

Data were automatically normalized to lowercase, stripped of any punctuation and automatically coded as correct or incorrect using a simple Python script. This script set a boolean 'isCorrect' variable to true if the target word was present in the

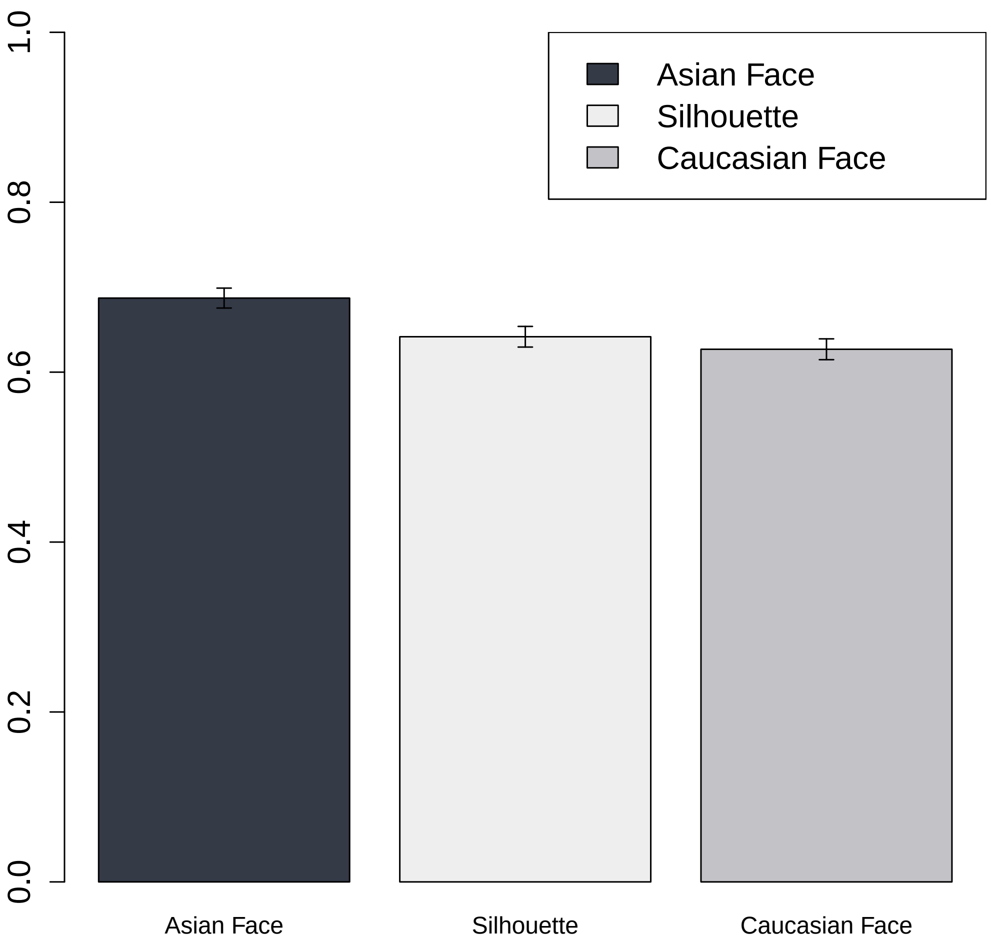

Figure 3.2: **All listeners: proportion correct target word responses for combined inexperienced and experienced listeners**

	Coef β	SE(β)	**z**	**p**
(Intercept)	2.06	0.28	7.3	**<.001**
Silhouette	−0.24	0.13	−1.9	0.0638
Caucasian Face	−0.35	0.13	−2.7	**.0064**
Low Predictability	−1.17	0.07	−15.7	**<.001**
experienced	−0.58	0.11	−5.2	**<.001**

Table 3.2: **Correct responses by Face, Predictability and Experience level**

text typed by the participant and false otherwise. These automated decisions were reviewed by a research assistant who was naive to the goals and design of the experi-

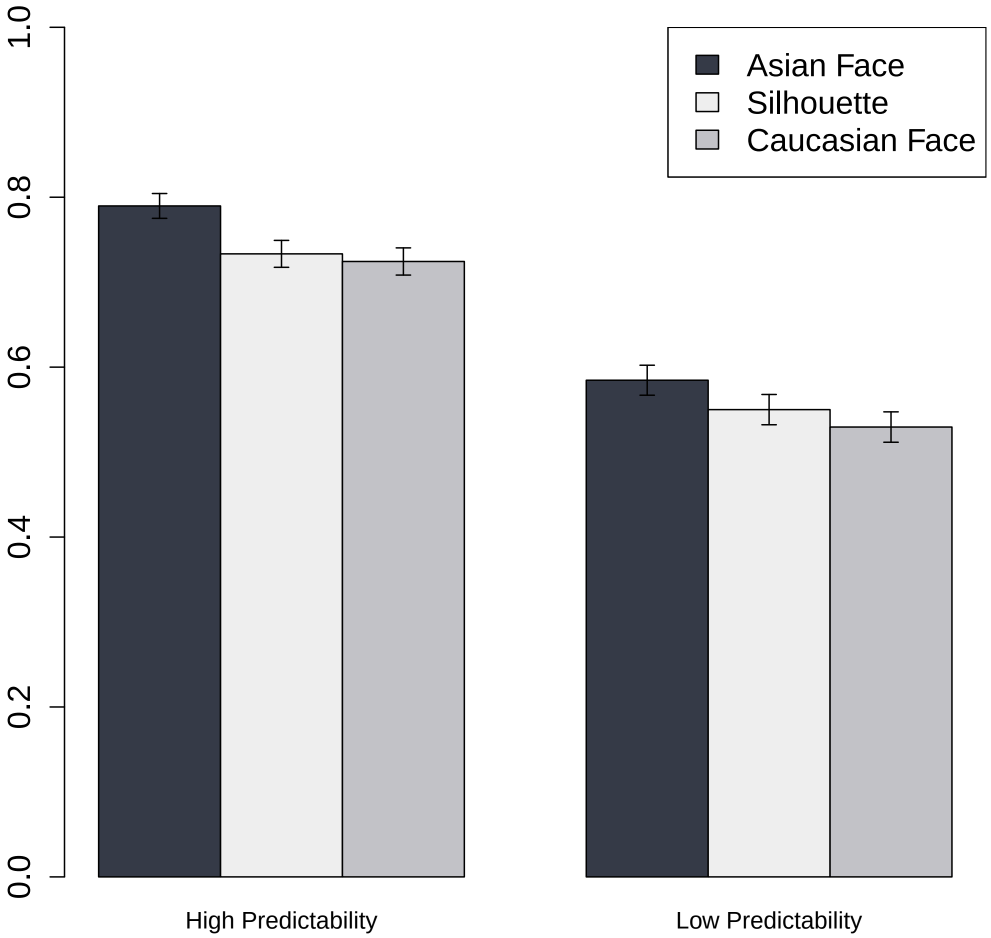

Figure 3.3: **All listeners: proportion correct target word responses for combined inexperienced and experienced listeners separated by Predictability**

ment. A small number of coding decisions were reversed for being mere typographical errors (e.g. “coffe” for *coffee* or “yello” for *yellow*). A response was coded as correct only if it contained the target (final) word in the sentence; a response satisfying this criterion could be otherwise blank or contain gibberish and still be ‘correct’. These coded transcription responses were then analyzed using the open source statistical package R 2.13.0 (R Development Core Team, 2011) and the packages `lme4` (Bates et al., 2011) and `languageR` (Baayen, 2008).

Figure 3.2 shows the proportion correct responses for all listeners in each Face condition pooled across both Experience level and sentence Predictability. Error bars in this figure represent standard error of the means, so while absence of overlap does not necessarily indicate significance, the presence of overlap virtually guarantees that the comparison in question is not significant. As this image suggests, there is a significant main effect of Face with transcription in the Asian condition significantly more accurate than transcription in the Caucasian condition. Table 3.2 reports the results of a linear mixed model analysis in which the Correct response variable is the dependent measure; Face, Predictability and Experience level are modeled as fixed effects; and Subject and Target word are random effects with random intercepts. With Asian face as the default reference level, the Caucasian face is significantly less accurate ($\beta = -0.35, p < .01$).

	Coef β	SE(β)	**z**	**p**
(Intercept)	1.38	0.28	5.0	**<.001**
inexperienced	−0.54	0.18	−3.0	**<0.01**
Silhouette	−0.12	0.21	−0.6	>0.56
Caucasian Face	−0.40	0.20	−2.0	**<0.05**
inexperienced:Silhouette	−0.14	0.26	−0.6	>0.57
inexperienced:Caucasian Face	0.12	0.25	0.5	>0.64

Table 3.3: **The interaction of Face and Experience in terms of correct responses by all listeners**

Figure 3.3 shows the proportion correct responses for the inexperienced listeners in each Face condition by sentence predictability. As this graph suggests, there is a significant main effect of Predictability ($\beta = -1.17, p < .001$). There is also a significant main effect of Experience ($\beta = -0.58, p < .001$). Experienced listeners were more accurate overall; however, as there is no interaction between Experience and Face, experienced listeners do not receive a greater or lesser benefit than inexperienced listeners when shown an Asian face and transcribing Chinese-accented English. In a linear mixed model with Correct response as the dependent variable, a single fixed effect interaction term of Face by Experience, and Subject and Target word included as random effects, the interaction is not significant (table 3.3). With default reference levels of ‘experienced’ for the Experience variable and ‘Asian Face’ for the Face variable, neither ‘Silhouette’ ($\beta = -0.1426, p = 0.58$) nor ‘Caucasian Face’ ($\beta = 0.1160, p = 0.64$) shows improved or diminished transcription accuracy.

The Silhouette condition does not differ significantly from Asian Face ($\beta = -0.24, p = 0.0638$), although this result would be significant at a higher $\alpha = 0.1$ level. Transcription accuracy in the Silhouette condition does not differ significantly from accuracy in the Caucasian Face condition when the reference level is switched to Silhouette and the model recalculated ($\beta = -0.11, p = 0.3821$).

	Coef β	SE(β)	**z**	**p**
(Intercept)	1.78	0.29	6.2	**<.001**
Low Predictability	−1.30	0.13	−9.9	**<.001**
Silhouette	−0.41	0.18	−2.3	**<.05**
Caucasian Face	−0.46	0.18	−2.6	**<.05**
Low Predictability:Silhouette	0.22	0.18	1.2	>0.22
Low Predictability:Caucasian Face	0.16	0.18	0.9	>0.37

Table 3.4: **The interaction of Face and Predictability in terms of correct responses by all listeners**

In these combined results at least, the prediction of an interaction between the variables Face and Predictability does not appear to have been upheld. Indeed, in a linear mixed model with Correct response as the dependent variable, a single fixed effect interaction term of Face by Predictability, and Subject and Target word included as random effects, the interaction is not significant (table 3.4). With default reference levels of ‘Asian Face’ for the Face variable and ‘High Predictability’ for the Predictability variable, neither ‘Silhouette’ ($\beta = 0.22, p > 0.22$) nor ‘Caucasian Face’ ($\beta = 0.16, p > 0.37$) shows improved or diminished transcription accuracy. In these combined results there appears to be no statistical difference between levels of

(ref. level: Asian Face:High Predictability	Coef β	SE(β)	**z**	**p**
(Intercept)	1.60	0.32	5.0	**<.001**
Silhouette	−0.51	0.20	−2.5	**<.05**
Caucasian Face	−0.41	0.21	−2.0	**<.05**
Low Predictability	−1.38	0.17	−8.3	**<.001**
Silhouette:Low Predictability	0.40	0.22	1.8	>0.1
Caucasian Face:Low Predictability	0.16	0.22	0.7	>0.5

Table 3.5: **Inexperienced listeners: fixed effects with coefficients and *p*-values for correct responses by Face and Predictability**

predictability. This observation is not upheld when the data are divided by Experience level. The meaningfulness of this finding will be discussed below.

All responses by inexperienced listeners that were coded as errors are included in Appendix A which reports errors in the High Predictability condition and Appendix B which reports errors in the Low Predictability condition. Appendix C and Appendix D present errors by experienced listeners. Errors in these appendices are sorted by target word and by frequency within target word.

3.2.9 Inexperienced Listener Results

It is worthwhile to divide the data by levels of the Experience condition and examine trends in the data for evidence of the influence of experience on transcription accuracy. Figure 3.4 shows the proportion correct responses by inexperienced listeners in each Face condition by sentence predictability. There is a 19.9% improvement in the High predictability condition at 71.6% correct versus 51.7% correct in the Low predictability condition.

As is evident from figure 3.4, the prediction of an interaction between the variables Face and Predictability was upheld for inexperienced listeners but in the opposite of the predicted direction. Rather than seeing, as predicted, greater benefit of information provided by the purported speaker's face in the Low predictability sentences,

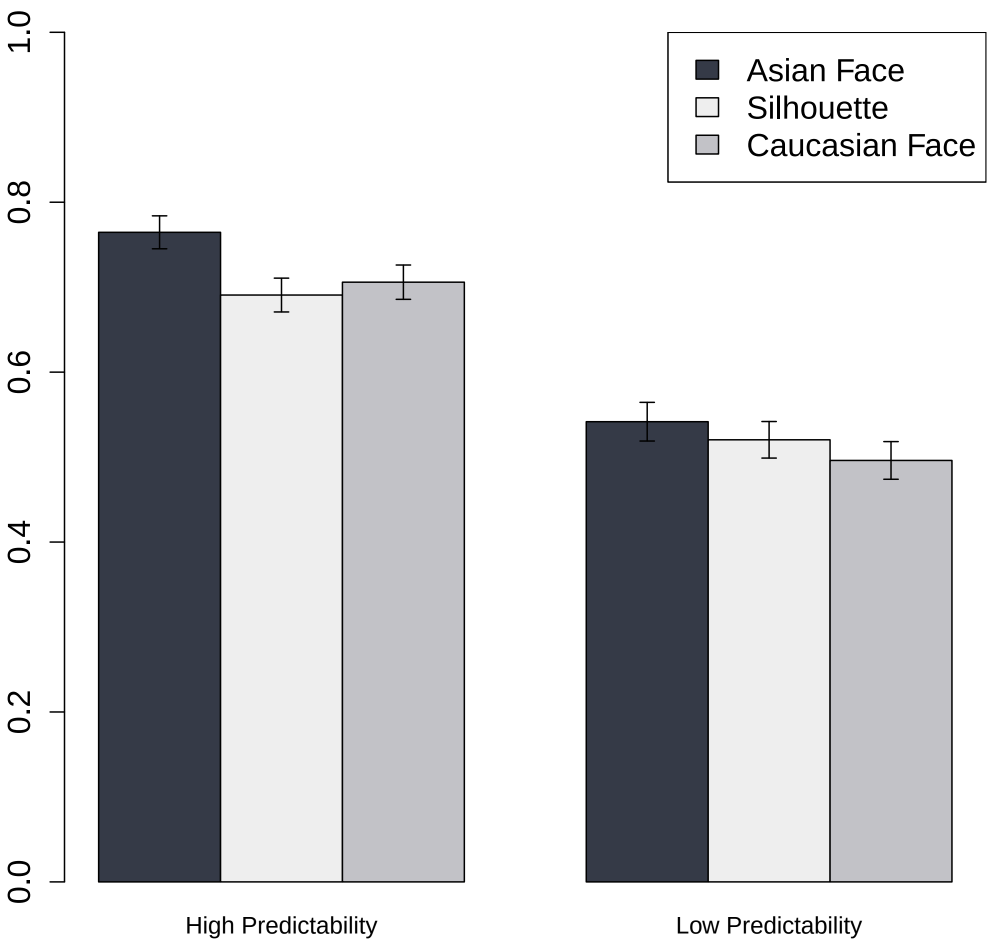

Figure 3.4: **Inexperienced Listeners: proportion correct target word responses**

there is essentially no benefit in this condition. Instead, listeners in the Asian face condition received the most benefit when semantic information made the target words highly predictable.

Reported in table 3.5, Face, Predictability and the interaction of Face and Predictability were included as fixed effects in a linear mixed model with Subject and Target word as random effects. Correct responses were, again, the dependent measure. There is no statistical difference between Asian face, Silhouette and Caucasian face for the Low predictability sentences in the Inexperienced listener condition. With

Asian Face and High predictability sentences as the reference levels, both Caucasian face ($\beta = -0.51, p < 0.05$) and Silhouette ($\beta = 0.41, p < 0.05$) are significant.

A simplified model with Predictability and the interaction term removed as fixed effects was run to test for a main effect of Face independent of other factors for inexperienced listeners. With a base reference level of Asian, the Silhouette condition is not significant ($\beta = -0.2727, p = 0.071$) but the difference between the Asian and Caucasian Face conditions just narrowly misses significance ($\beta = -0.296, p = 0.0526$). The trends in this pattern differ from the performance of Experienced listeners reported in the following section.

3.2.10 Experienced Listener Results

Figure 3.5 shows proportion correct responses in each Face condition by sentence predictability. There is an 18.7% difference between proportion correct responses on the High predictability sentences at 80.6% correct and Low predictability sentences at 61.9% correct. While overall accuracy is higher, the percentage improvement for the High predictability condition is nearly identical to the 19.9% improvement shown by inexperienced listeners.

As reported above, overall transcription performance is significantly better for experienced than inexperienced participants. One might argue that this overall difference is due not to the different experience levels of the listener populations but to some difference in quality between University of Michigan and UC Berkeley undergraduates. These schools are quite similar academically so this explanation seems less plausible than the experienced/inexperienced distinction, but it is a possible explanation that has not been controlled for in the design.

Once again, listeners received the most benefit when semantic information made the target words highly predictable. Subject and Target word were included as random effects in a generalized linear mixed model with binomial errors and a logit link

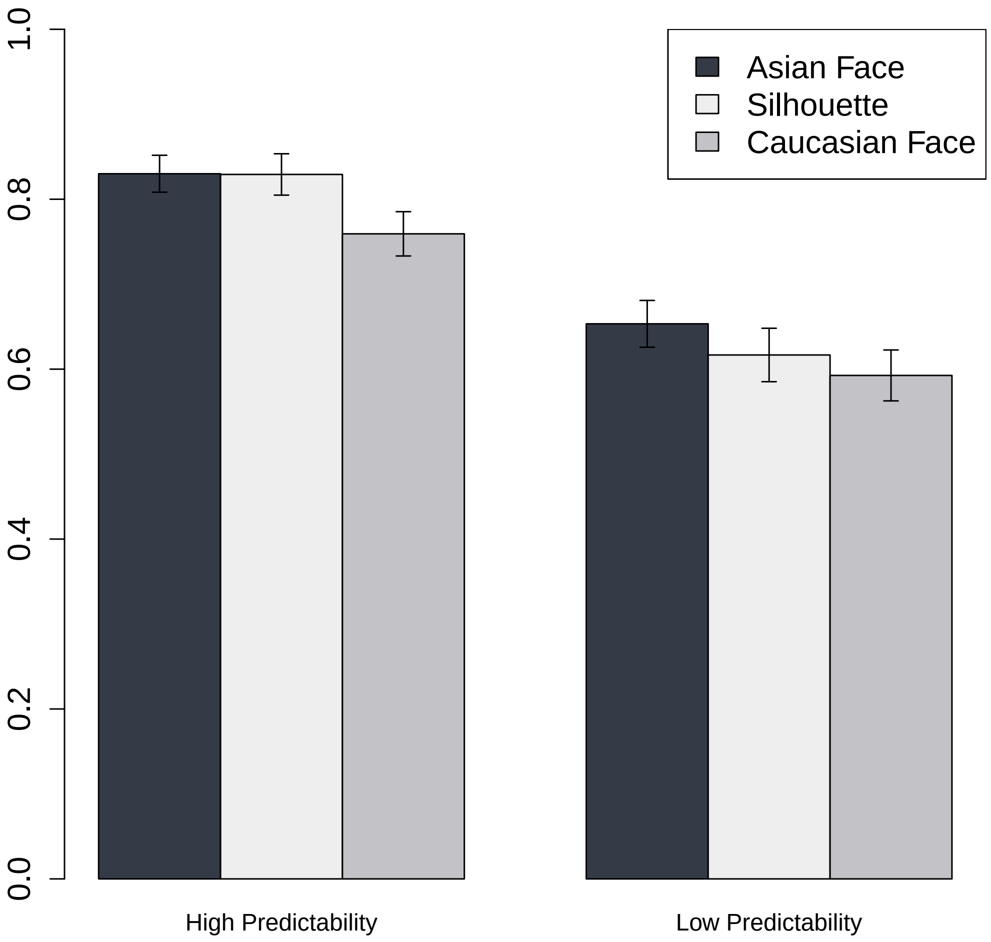

Figure 3.5: **Experienced Listeners: proportion correct target word responses**

	Coef β	SE(β)	**z**	**p**
(Intercept)	2.03	0.30	6.7	**<.001**
Silhouette	−0.01	0.29	0.0	0.9759
Caucasian Face	−0.54	0.27	−2.0	0.0501
Low Predictability	−1.20	0.22	−5.5	**<.001**
Silhouette:Low Predictability	−0.20	0.32	−0.6	0.5430
Caucasian Face:Low Predictability	0.20	0.30	0.7	0.5088

Table 3.6: **Experienced Listeners: fixed effects with coefficients and** p**-values for correct responses by Face and Predictability**

function. Face, Predictability and the interaction of Face and Predictability were once again included as fixed effects in the model. Table 3.6 reports output for the generalized linear model. There is once again no statistical difference between Asian face, silhouette, and Caucasian face for the Low predictability sentences. With Asian Face in the High predictability sentences as the reference level, neither Caucasian face ($\beta = -0.536, p = 0.0501$) nor Silhouette ($\beta = -0.009, p = 0.9759$) is significant at the p < 0.05 level. Caucasian face just misses significance and perhaps a larger number of subjects would have provided the signal required to discern the real difference in the levels, but this is merely conjecture.

There is a main effect of Face in a generalized linear model with only Face as a fixed effect ($\beta = 0.0428, p = 0.0428$).

3.2.11 Discussion

To the extent that all other possible factors were successfully held constant in the experiment, it is reasonable to infer that a facilitatory effect of purported speaker face has occurred at the level of the processing of the speech stream after it has reached the listeners' ears. Listeners, even inexperienced listeners, show improved performance on this transcription task when the face they are shown provides socioindexical information consistent with the voice they are listening to. However, it is not at all clear that the facilitation seen in this experiment is due to the influence of social knowledge on the processing of fine phonetic detail.

The fact that the silhouette condition has patterned with the Caucasian face for inexperienced listeners and is significantly different from the Asian face, at least in the high predictability condition, suggests that the listeners' default expectations about speaker identity and the socioindexical information conveyed by the Caucasian face are highly similar. That interpretation is perhaps unsurprising, but the corollary of this observation is that, at least when the face is Caucasian and the voice is authentic Chinese-accented English, there is apparently no additional inhibitory effect of mismatched visual and auditory socioindexical information. I will return to this surprising result in more detail in Chapter IV.

It is fascinating to compare listeners' performance in the silhouette condition of the experienced and inexperienced listeners. Though, with the reference level of the Face variable changed to 'Silhouette', silhouette is not significantly different from Caucasian face for experienced listeners ($\beta = -0.53, p = 0.0696$), this condition does clearly appear to cluster with the Asian face condition –in contrast to the Silhouette results in the Inexperienced condition, which clustered with the Caucasian face condition and were significantly different from the Asian face. It may well be that experienced listeners' default expectations about speaker identity, when listening to Chinese-accented English, are highly similar to those expectations established by the presentation of the purported Chinese speaker's face. We must be cautious about this conclusion, however, as these listeners knew that they had been recruited because they were Chinese Americans with Heritage Mandarin experience[2]; this could easily have influenced their default assumptions about the identity of the speaker in the silhouette condition.

I am personally most surprised by the finding that the visual stimulus is only significantly helpful in the High predictability sentences for either experienced or inexperienced groups of listeners. Although this significant interaction disappears when results are combined due to the larger number of observations in the combined model; this suggests that the interaction of Face and Predictability within the Experienced and Inexperienced groups of listeners is weak and may be a spurious result due to low sample sizes.

However, it does seem clear that the magnitude of the improvement listeners experience when seeing the Asian face and hearing Chinese-accented English tends

[2] Given this knowledge it is perhaps even more impressive that the Caucasian face manipulation appears to differ from the Asian face condition as clearly as it does.

to be greater for High Predictability sentences than for Low Predictability sentences. Perhaps socioindexical information is most useful when it can reinforce, and possibly clarify, semantic information. It would be interesting to replicate this experiment using a lower signal-to-noise ratio to see if, when the acoustic signal is clearer, the effect of socioindexical expectation is more visible on the low semantic predictability sentences. My prediction is that this is precisely what would happen.

Contrary to the overwhelming view in the sociophonetics perception literature and to my own predictions, it does not appear to be the case that inexperienced listeners are making use of socioindexical knowledge to alter their specific predictions about fine phonetic detail. Few, if any, of the most common mishearings by either experienced or inexperienced listeners are consistent with the hypothesis that listeners use experience with or stereotypes of Chinese-accented English when processing fine phonetic detail. In both High and Low predictability contexts, for example, listeners frequently heard *sport* as *spot* (the most common mishearing overall). Sport, in the stimulus recordings, was produced without a clear post-vocalic consonantal [ɹ]. The absence of post-vocalic [ɹ] is a strongly stereotypical feature of Chinese-accented English. If listeners were using either their experience with Chinese *or* their stereotypes of Chinese to anticipate accented speech we would expect them to reconstruct this missing [ɹ]. This is especially true given that the speaker, as demonstrated in Chapter II, produces a fairly rhoticized vowel, despite the missing consonantal articulation, in this token. Further analyses of specific transcription errors and how well they are, or are not, predicted by actor imitations of Chinese-accented English are part of a subsequent project and outside the scope of this dissertation.

3.3 General Discussion

Experienced listeners are overall better at the task than inexperienced listeners. In a sense, this difference replicates the usual finding in the sociophonetic perception literature –it is clear that listeners' experience with an accent affords them greater facility when transcribing that accent in noise. However this difference is not, in itself, evidence for an episodic trace theory of the lexicon. Furthermore, the usefulness of socioindexical knowledge to inexperienced listeners appears to be of roughly the same magnitude as it is to experienced listeners. This socioindexical knowledge is also disproportionately more helpful in High predictability sentences than Low and does not seem to result in an abundance of mishearings consistent with an accented set of lexical or prelexical linguistic expectations.

Returning to the discussion of Staum Casasanto's work from Chapter I, if the bottom-up mechanism she proposes for the influence of socioindexical knowledge on speech perception were at work here, or the mechanism proposed in Johnson (2006), we would expect to see facilitation only for experienced listeners and, for that population in particular, mishearings showing compensation for/expectation of Chinese-accented productions. The results of the present experiments, then, are inconsistent with these exemplar models in which socioindexical knowledge preactivates stored exemplars consistent with that socioindexical label.

A useful follow-up experiment will be an AXB discrimination experiment in which I manipulate just as I have here, listener expectations about the identity of the speaker. It will be interesting to discover if the socioindexical manipulation can, in such a task, shift category boundaries in the directions predicted by real and stereotypical features of Chinese-accented speakers. The pattern of errors in the present transcription task suggests that this will not be the case.

Finally, one conclusion we can definitively reach based on the results of these experiments is that Rubin and Lippi-Green's interpretation of the results in Rubin (1992) is not consistent with these findings. It is not the case that monolingual English speakers presented with a purportedly Asian-faced speaker tune out that speaker or refuse to uphold their end of the communicative burden. In the present experiment, just as in the Rubin study, when the face provided the listener with informative information about the identity of the speaker, performance was improved. Conversely, when the displayed face provided the listener with misleading information about the identity of the speaker, again *just* as in Rubin, performance was lower.

Given the improved performance of inexperienced listeners on this task, it does appear to be the case that listener stereotypes of Chinese accented English play a role in speech perception, but that role is demonstrably not negative.

What remains to be explored is when exactly, during speech perception, this influence of socioindexical expectation is detectable. The time course of the influence of socioindexical knowledge on speech perception is examined in Chapter IV.

CHAPTER IV

The Influence of Socioindexical Expectation on Speech Perception: Evidence from Eye-Tracking

The experiment presented in Chapter III investigated the influence of manipulated socioindexical expectations on the transcription of Chinese-accented speech in noise. This experiment showed that, in the high predictability condition, Asian face condition listeners were more accurate than Caucasian face listeners on a sentence-in-noise transcription task. It is tempting to interpret an experimental result like this one as evidence about listeners' use of fine phonetic detail during speech perception. However this task can really only illuminate the *outcomes* of speech perception and word recognition processes and not the time course of the perceptual mechanisms. We can infer from these outcomes certain properties that the speech processing system must have: listeners can use visual information to alter their interpretation of acoustic information when arriving at a speech percept, manipulating visual representations of racial information can alter the interpretation of acoustic information, recognition improves when visual cues to speaker identity match auditory cues to speaker identity, etc..

What is not clear from performance on the transcription task is the time course of listeners' use of socioindexical information. It could be, for example, that experienced and inexperienced listeners arrive at similar outcomes on the transcription task but do so by entirely different means. A related but, I think, somewhat more fundamental question is how the processing of an identical auditory stimulus differs across listeners in the different Face conditions.

We have some evidence from patterns of mishearings that socioindexical knowledge does not exert a bottom-up influence on perception by shifting listeners' expectations of fine phonetic detail. Without access to the time course of these mishearings, though, it is impossible to tell whether the auditory stimuli really were processed identically across Face conditions or whether listeners in the Asian Face condition considered, for example, 'sport' as a strong candidate even when they ultimately typed 'spot'. It may even be the case that listeners considered these forms even when they ultimately typed nothing at all.

The findings of the transcription task can not eliminate the possibility that social cues exert a bottom up influence on perception in the form of shifted phonological expectations, an altered parsing strategy or even heightened attention. Alternatively, listeners in the various socioindexical conditions may be processing acoustic information identically in which case social knowledge may exert a top down influence by suppressing or promoting particular forms.

Furthermore, the finding, suggested by the transcription results, that a mismatch between visual and auditory cues to speaker identity has no apparent inhibitory effect is sufficiently surprising to merit further investigation. It may be the case, as Lippi-Green (1997) implies, that, regardless of the socioindexical information in the voice of the speaker, one should *not* expect an inhibitory effect when the visual representation of the speaker is Caucasian. That is, the inhibitory effect in the Rubin study is the result of racism or anti-ethnic bias on the part of the listener as a result of seeing an Asian face and not the voice/face mismatch. Or perhaps the mismatch between voice and face actually does affect listeners' processing of speech-in-noise during the transcription task and evidence of this processing difficulty is masked by a separate benefit for seeing a Caucasian face. With only the eventual outcome to work from, the results are open to a large number of interpretations.

To better understand speech perception, I believe it is necessary to explore the processing underlying the outcomes in experiments like mine. In the present chapter I describe a set of eye-tracking experiments using a visual world paradigm. This paradigm was originally developed to address effects such as lexical frequency or neighborhood density, precisely the class of issues I am now raising about the processing of socioindexical information. Eye-tracking is an established tool in psycholinguistics for studying the unfolding of lexical decision processes over time, from the arrival of acoustic information to word recognition (Eberhard et al., 1995; Magnuson et al., 1999; Allopenna et al., 1998; Dahan et al., 2001; Beddor et al., 2009).

One might also study on-line speech processing using imaging techniques such as event related potentials (ERP), which offers high temporal accuracy of brain activation, or functional magnetic resonance imaging (fMRI), which offers high spatial accuracy of brain imaging. One benefit eye-tracking offers over either of these imaging technologies is a measurable behavioral link between stimulus and processing. Eye-tracking has proven to be a powerful tool for the investigation of speech processing because it measures a behavioral response that can be straightforwardly linked with both processing and with the listeners' objective when completing a task.

In the visual world paradigm, a listener is presented with a set of representations of objects on a computer screen and asked to listen to an acoustic stimulus. The listeners' eye movements are then tracked and recorded with high speed cameras so that saccades (fast eye movements) and fixations (periods of relative stability in the gaze) can later be analyzed. Eye movements, measured in this way, directly indicate how a listener's visual attention is allocated and therefore only indirectly indicate how the acoustic stimulus is being processed (Huettig et al., 2011). However, the link between processing and eye movements has proven to be robust and informative

(Huettig et al., 2011; Allopenna et al., 1998).

In addition to linguistic details, the visual world paradigm is sensitive to, for example, the imageability of the different objects on screen. A more visually engaging image is likely to draw more attention than a less visually engaging image –regardless of the acoustic stimulus. In the eye-tracking studies presented here, all aspects of the actual stimulus presentation are held constant across trials. The manipulation of socioindexical expectation happens prior to the presentation of the visual world or of the acoustic stimulus so that any differences in attention are presumably attributable to the cognitive state of the listener and not to, for example, the attractiveness of the purported speaker, the visual appeal of a particular image, mismatches in imageability between target word and competitor, or similar issues.

4.1 Experiment 3: Time Course

The present experiment is, in spirit, a replication of Rubin's (1992) experiment 1 using eye-tracking to investigate the time course of listeners' processing while manipulating socioindexical expectations. Rubin's participants listened to tape-recorded lectures while viewing an image of either an Asian or Caucasian purported graduate student instructor. After listening, participants completed a cloze test in which every seventh word of the mini-lecture had been elided and needed to be supplied by the participant. The participants then responded to homophyly, ethnicity, and accent ratings on a set of seven point Likert ratings scales.

Rubin found a main effect of Face on perception of accent. Listeners in the Asian-faced instructor condition were significantly more likely to rate the same voice as more accented: 3.77 out of 7 versus 2.75 in the Caucasian face condition. Though not statistically significant, these listeners also tended to have lower scores on the cloze test of comprehension: scoring, on average, 2.01 points lower than listeners in the Caucasian face condition on humanities lectures and 5.19 points lower on science lectures.

Lippi-Green (1997) describes these results in the most stirring terms. Listeners in the Asian face condition are, according to Lippi-Green, shirking their communicative responsibility. Following Rubin's own interpretation, Lippi-Green claims Rubin's findings indicate that "preconceptions and fear are strong enough motivators to cause students to construct imaginary accents, and fictional communicative breakdowns." (Lippi-Green, 1997, p. 128). The assumption underlying these claims is that Rubin's results indicate reduced attention on the part of Asian face condition listeners. Due to racial bias these listeners are simply not attending to the acoustic signal as closely as those in the Caucasian face condition. Given Rubin's task, though, lack of attention must be inferred from the outcomes rather than observed. Whether listeners actually show reduced attention is an empirical question and one that eye-tracking is particularly well suited to investigating.

In the present experiment, Rubin's cloze task is replaced with a visual world experiment intended to directly assess listeners' attention to and processing of auditory stimuli in an inverse matched guise manipulation. Replacement of missing words in a text is replaced by fixations to particular a target image when presented alongside an image representing a phonologically similar distracter word. This task is, admittedly, less ecological than Rubin's, but I have traded naturalness for precise instrumentation.

4.2 Methods

4.2.1 Stimuli

4.2.1.1 Auditory Stimuli

The auditory stimuli for this experiment were natural speech recordings of a native speaker of Standard American English (SAE) from San Diego, California who has extensive training in phonetics. She read each word 5 times in sequence so that the fifth, utterance final, production could be tokenized. The tokenized recordings were then trimmed of leading and trailing silence and amplitude normalized.

Though at the time of stimulus construction there was no theoretical reason to presume that the identity of the speaker might be of further importance, it should be noted that she is a Heritage speaker of Mandarin Chinese. This was a methodological oversight. In addition to overhearing Mandarin conversation between her parents (see Au et al., 2002) and being addressed directly in Mandarin by her maternal grandparents, she received explicit training in Mandarin Chinese outside the home and retains receptive and basic speaking competence into adulthood. This speaker has no discernible Chinese accent and has provided stimuli for numerous speech experiments, but the findings of Newman and Wu (2011) force me to consider the possibility that, as Whalen (1984) eloquently posited for subcategorical coarticulatory information, listeners may be sensitive to indexical markers in the speech stream that experimenters cannot hear directly. It is an open question whether the results would have looked somewhat different with a monolingual English speaker, inexperienced with Mandarin.

4.2.1.2 Visual Stimuli

One of two facial images was presented to listeners to establish the expectation of an accent; these faces are from a database collected by Minear and Park (2004) and are shown in figure 4.1. Kennedy et al. (2009) normed the Minear & Park faces with Likert scales for perceived age, familiarity, mood, memorability, and picture quality and these pictures were chosen for their similarity on those dimensions. Each listener saw only one of the two images (between-subjects design). The image was displayed prior to the presentation of each trial.

Two images, one target and one competitor, were displayed on-screen during each

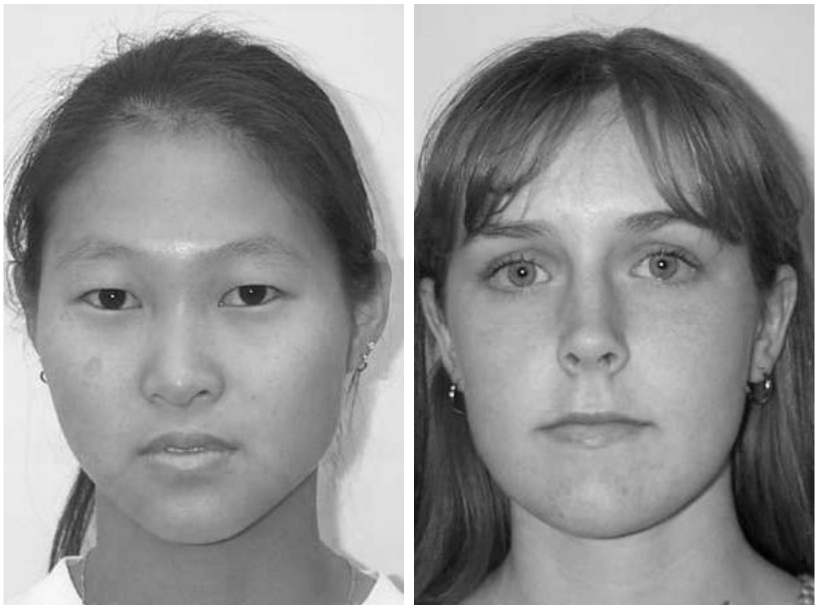

Figure 4.1: **Minear and Park (2004) faces used in the eye-tracking experiment**

trial. Many of these images were developed for Beddor et al. (2009), one was drawn by the author and several were found via web search. All stimulus images used in this experiment are shown in Appendix E.

4.2.2 Apparatus

The experiment used an Eyelink II head-mounted binocular eye tracking device with a pair of headgear-mounted 500Hz cameras configured for pupil tracking without corneal reflection. Cameras were focused on the participant's pupils followed by a 9-point calibration procedure to independently establish the degree of eye movement required for each of the participant's eyes to locate and fixate the periphery of the 17" 1024x768 CRT display.

The eye tracking system in Dr. Boland's psycholinguistics lab uses a grid of four infrared emitting diodes to locate the headgear with respect to the display which is itself mounted on a height-adjustable table. Care was taken during the calibration procedure to ensure that each participant was seated comfortably, facing forward, and centered horizontally and vertically with respect to the screen. Participants were seated approximately 24 inches from the presentation screen. Fixations were recorded only for each participant's dominant eye as identified by the SR Research system.

Auditory stimuli were presented over AKG K271 mkII headphones in a quiet, but not sound-attenuated, experiment room; volume was set at a comfortable listening level. The experimenter was present in the room for the duration of the experiment to monitor and calibrate from a separate experimenter's computer. This computer and the experimenter were visually and acoustically separated from the participant by a free-standing, sound-absorbing cubicle wall. Both auditory and visual stimuli were presented using SR Research Experiment Builder software version 1.6.121.

4.2.3 Procedure

Prior to their arrival, participants were randomly assigned to one of the two Face conditions: Asian or Caucasian. Participants completed a brief familiarization task in which they were presented with the series of stimulus images and their associated labels. Participants were asked to say the label out loud, describe to the experimenter how they imagine the label might be associated with the image and then repeat the label again before the experimenter advanced to the next image/label pair. The repetition of the image name was intended to help the participant learn the link between word and image. Describing the link between the word and the image was intended to reinforce that such a link might exist and, again, to help the participant remember the label during the eye-tracking experiment. Images were then presented without labels and participants were asked to name the image from memory. This naming procedure was repeated until all images could be named without error; no participant needed more than two iterations on the naming task to achieve 100% recall.

Participants read the following instructions presented on the computer screen:

> Each trial will begin with the calibration target (a small circle) in the center of the screen. This serves as a check that the eye-tracker is still tracking your eye movements accurately. Once you look at it, this circle should disappear and will be replaced by a photo of your speaker.
>
> Next, two drawings will be displayed. We will first ask you to look at the drawings. Then a yellow cross will appear between them and we will ask you to look at it. The cross will disappear when we ask you to look at one of the drawings. Try not to move your eyes before you hear the picture label, but look at the appropriate drawing as soon as you think you know what the word is.
>
> The first 5 trials will be practice.

There were five practice trials to familiarize participants with the procedure, find a comfortable listening level and address any discomfort associated with the apparatus.

Before each trial, the experimenter completed a drift correction procedure which served not only to maintain calibration of the eye tracker but also to draw participants' attention to the center of the screen. At the completion of the drift correction procedure, and prior to the apparent beginning of the trial itself, participants were shown the image of their purported speaker for 2000ms. The selection of 2000ms was somewhat arbitrary, but was intended to allow participants sufficient time to see the face, extract racial information, and activate any higher level social stereotypes or experience the listener may have to draw on given that racial information. In an ERP study of the time course of social perception, Ito et al. (2004) found that, on average, participants identified faces within approximately 160ms and were able to make in-group/out-group determinations by as early as 250ms after the presentation of a face. Participants in the present eye-tracking experiment clearly had enough time to evaluate the face of their purported speaker. It is an open question whether they were given too *much* time to ponder the face and how a shorter stimulus presentation might have affected the results.

The face image was replaced by the target and competitor images and a fixation cross at the center of the screen. Listeners heard a recording of the experimenter's voice instruct them, "Look at the pictures." After 3750ms to view the drawings, recordings of the experimenter's voice gave the instructions, "fixate cross. Now look at...". These instructions were followed by the disappearance of the fixation cross and the simultaneous presentation of the stimulus recording of the target word. Target and competitor images remained on-screen for 1300ms after the onset of the auditory stimulus. Saccades and fixations to the target or competitor images were recorded for an interest period of 1000ms.

There were two enforced breaks after the practice items and after 45 trials.

4.2.4 Participants

Eighteen undergraduate students from the University of Michigan Introductory Psychology subject pool participated for partial course credit. Participants had no known hearing problems and reported no knowledge of Chinese or Chinese-accented English. Two participants completed the familiarization task but were unable to participate due to difficulty calibrating the eye-tracker. Data for one participant was lost after collection due to experimenter error. Data for the remaining 15 participants are reported here: seven in the Asian face condition, eight in the Caucasian face condition. Listeners completed a brief language history questionnaire after completing the eye-tracking task.

4.2.5 Predictions

If Rubin and particularly Lippi-Green's interpretation of Rubin (1992) are correct about the listener's role in Rubin's listening task then we should see a number of measurable behavioral results. First, we would expect listeners to indicate hearing

Chinese-accented English when shown an Asian face and asked about the accent of their speaker. Asian condition listeners should have longer first fixation latency, on average, than Caucasian condition listeners. Longer first fixation latencies could indicate that the listener is not making use of phonetic information as it arrives or is otherwise less engaged in the task. Lack of attention would also predict lower overall accuracy if the listener truly is shirking their end of the communicative burden and not attending to the auditory stimulus because of the speaker's face. Finally, Asian condition listeners, in this model, should make fewer, or shorter, overall fixations to the target image; this would be an indication, not merely of poor performance on the task, but of refusal (at some level, not necessarily as agentive as Lippi-Green hypothesizes) to pay attention.

Another possible interpretation of the transcription task results is that listeners experience a degradation in performance when the perceived indexical properties of the purported speaker's face fail to match the perceived indexical properties of the auditory stimulus. Exemplar theories in which social knowledge serves to pre-activate indexically appropriate exemplars would make this same prediction (Johnson, 2006, e.g.). Assuming that Asian faces are linked to Chinese (or other Asian non-native) accented English, exemplar models would, like Lippi-Green, predict longer first fixation latencies in the Asian Face condition. I believe these approaches also predict lower overall accuracy on the task. The key difference between these theories and Rubin/Lippi-Green is that there is no prediction of reduced overall attention (measurable as fewer overall fixations and lower overall time spent fixating the images). A difference on this dimension would serve to tease apart these two classes of prediction.

4.3 Results

At the end of each session, as the eye-tracking apparatus and headphones were being removed, the experimenter asked, “did your speaker have an accent?”. Caucasian

Face listeners unanimously reported that the speaker did not have an accent. Asian Face listeners unanimously reported that the speaker *did* have an accent and many of these listeners offered additional subjective responses such as, "Yes, but she's doing very well." or "Yes, but she's nowhere near as bad as my physics [graduate student instructor]!". These listeners seemed unaware that the speaker had no Chinese accent and, indeed, many seemed quite surprised when the experimenter revealed the deception that had taken place.

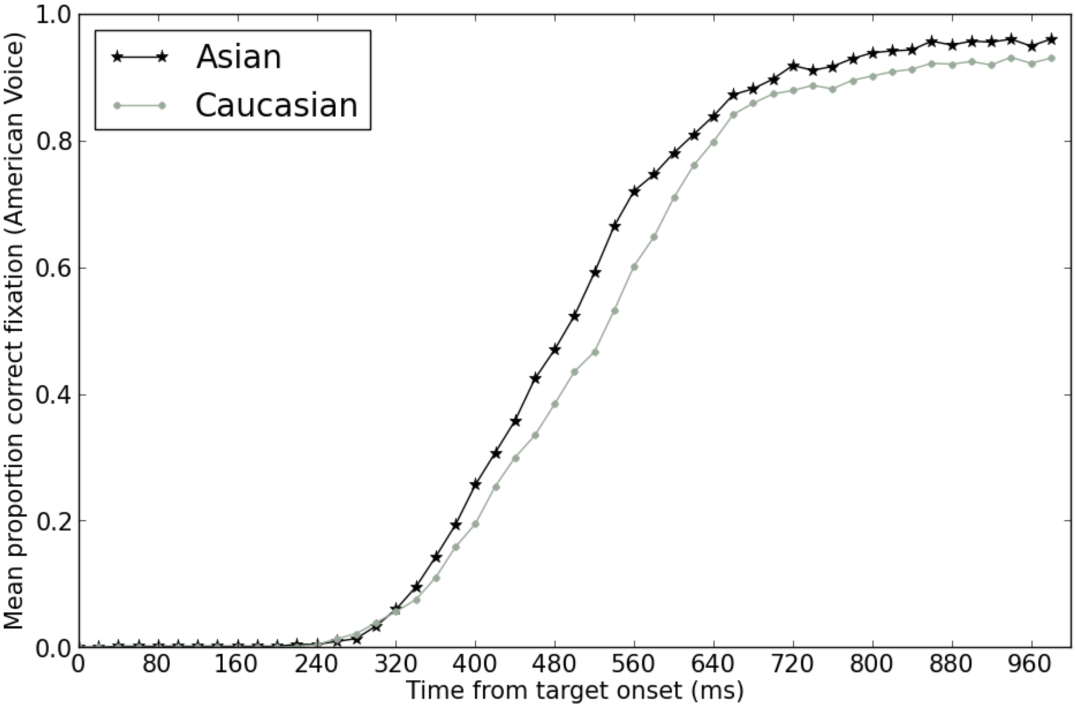

Figure 4.2: **Time Course of Mean Target Fixations**

Listeners in the Asian face condition show no inhibition for the mismatch of Face and accent. Although listeners need not be consciously aware of a mismatch to show its influence in processing time, it is interesting that they report hearing a Chinese accent and show no apparent inhibition in reaction time. Indeed, as both figures 4.2 and 4.3 show, listeners in the Asian Face condition appear to have been faster overall

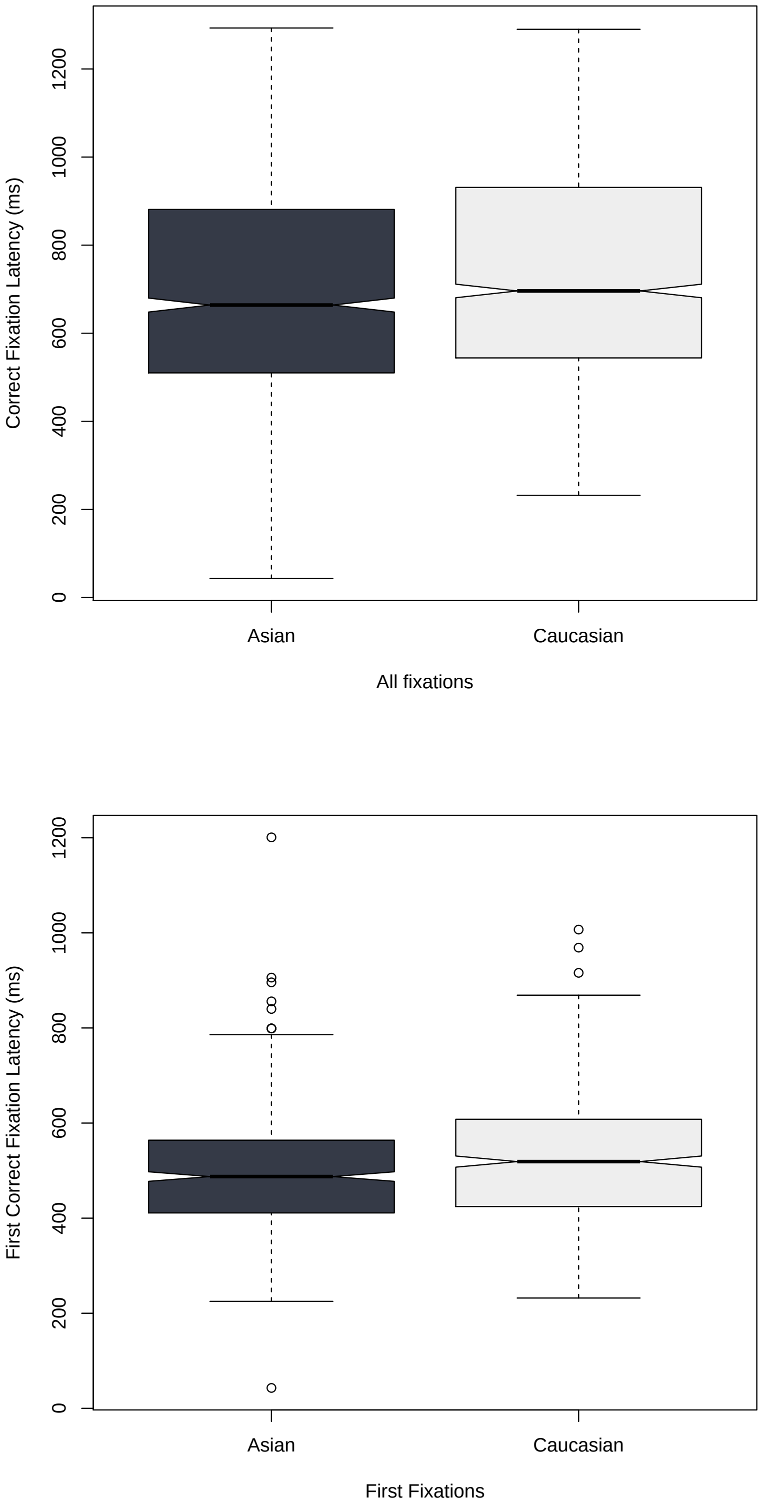

Figure 4.3: **Mean fixation latency by Face condition** top: full trial, bottom: first fixations only

to look at the target image and, as shown in the bottom image in figure 4.3, made faster first correct fixations as well. Listeners in the Asian Face condition also appear to have made a higher proportion of first correct fixations (figure 4.4, bottom image); a trend which held for correct fixations across the course of the trial (figure 4.4). The time course plotted in figure 4.2 for averaged fixation latencies for all subjects and all trials suggests that Asian Face listeners establish this dominance approximately 340ms after the onset of the auditory stimulus and maintain it throughout the trial.

However, the statistics do not support this apparent trend in fixation latencies. As might be immediately obvious from figure 4.5, there is sufficient between-subject variation in this between-subjects design to preclude significance.

When both Subject and Target word/image were included as random effects in a linear mixed model in which the two-level factor Face was the lone fixed effect and first fixation time was the dependent measure, the difference between first fixation latencies is not significant ($\beta = 35.81, p = 0.2513$). However, the number of subjects is small and given the apparent, if noisy, trend in figures 4.5 and 4.6, I suspect a larger sample might make it possible to detect a significant trend.

The statistics similarly do not support the apparent trend of higher accuracy in the Asian face condition implied by figure 4.4. Subject and Target word were included as random effects in a generalized linear mixed model with binomial errors and a logit link function. Face was included as a fixed effect. With a random intercept for Subject, the difference between proportion correct fixations in the Asian/Caucasian Face levels is not significant ($\beta = -0.5645, p = 0.17$).

The results in figure 4.7 suggest that listeners in the Asian face condition spent a larger proportion of the trial time fixating the target image. Plots of image dwell time for each image are included in Appendix E. Here the statistics do support the apparent trend in the visualized data. The difference in overall dwell time is significant with Subject and Target item included as random effects in a linear mixed

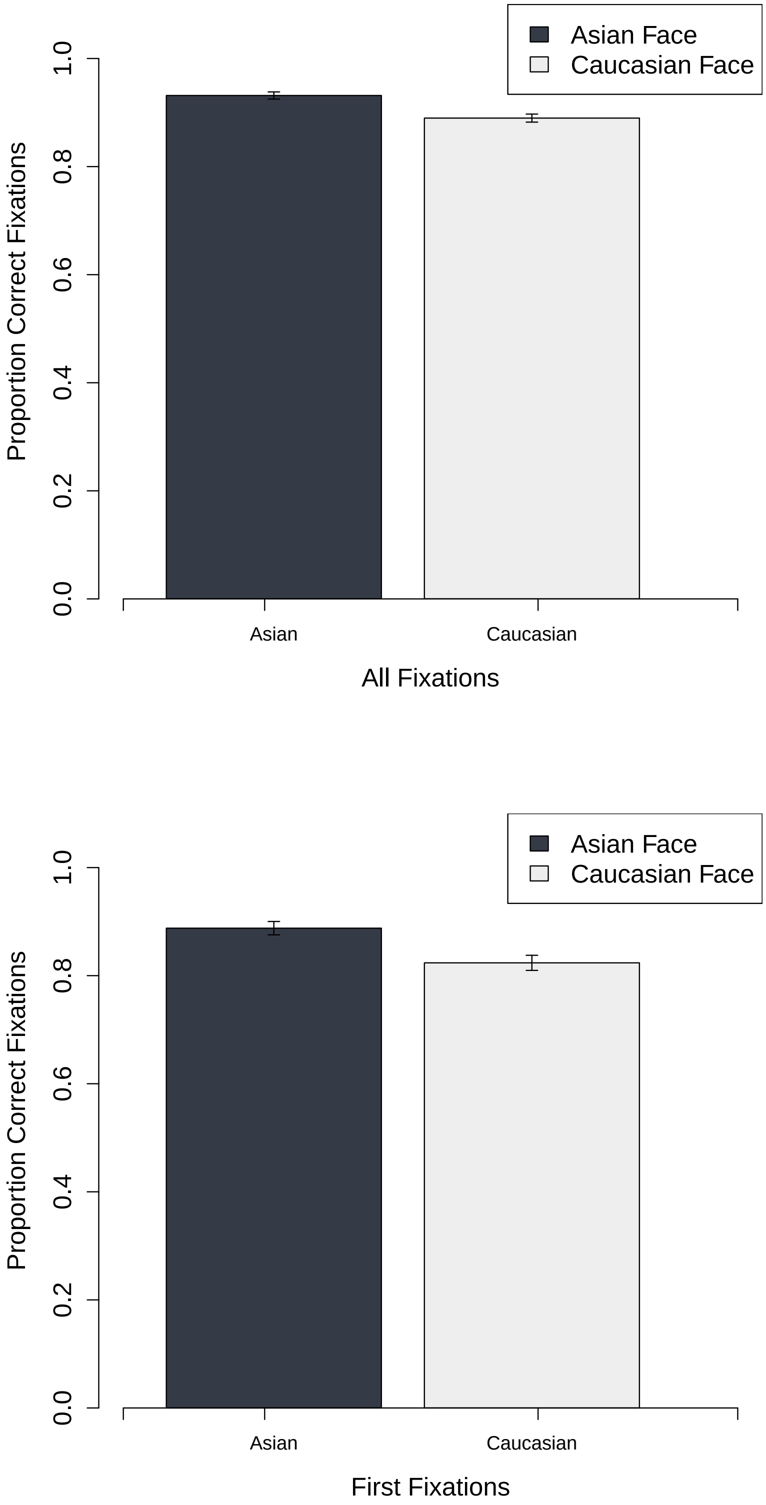

Figure 4.4: **Proportion correct fixations by Face condition: top: full trial, bottom: first fixations only**

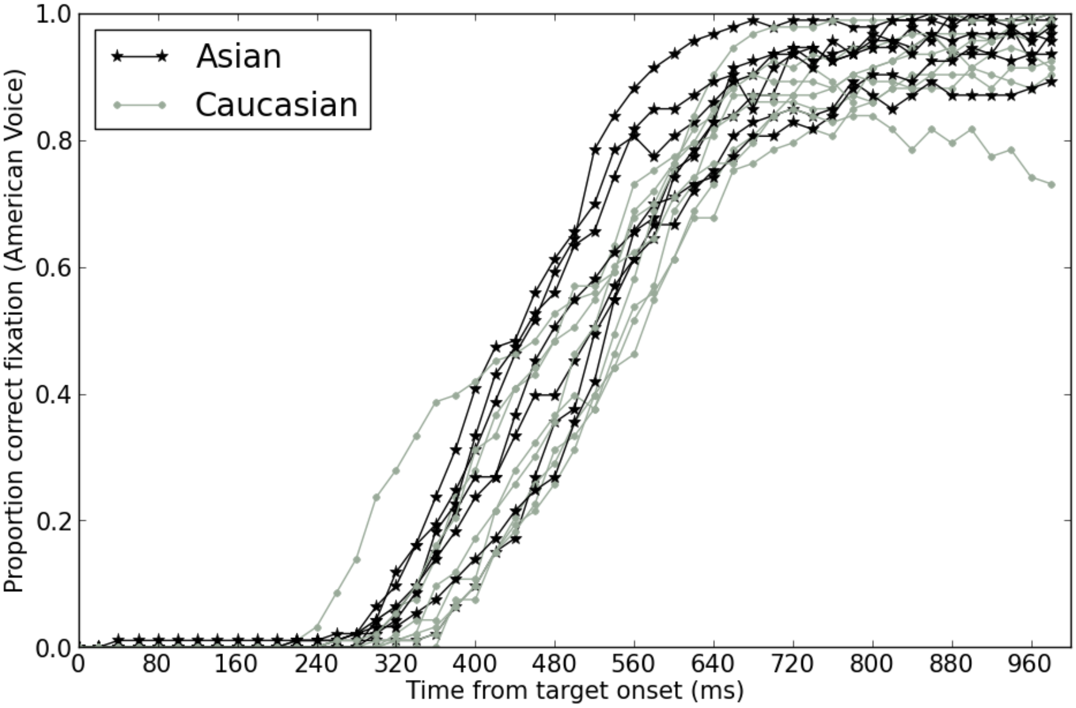

Figure 4.5: **Time Course of Individual Subject Mean Fixations** Line type indicates face condition

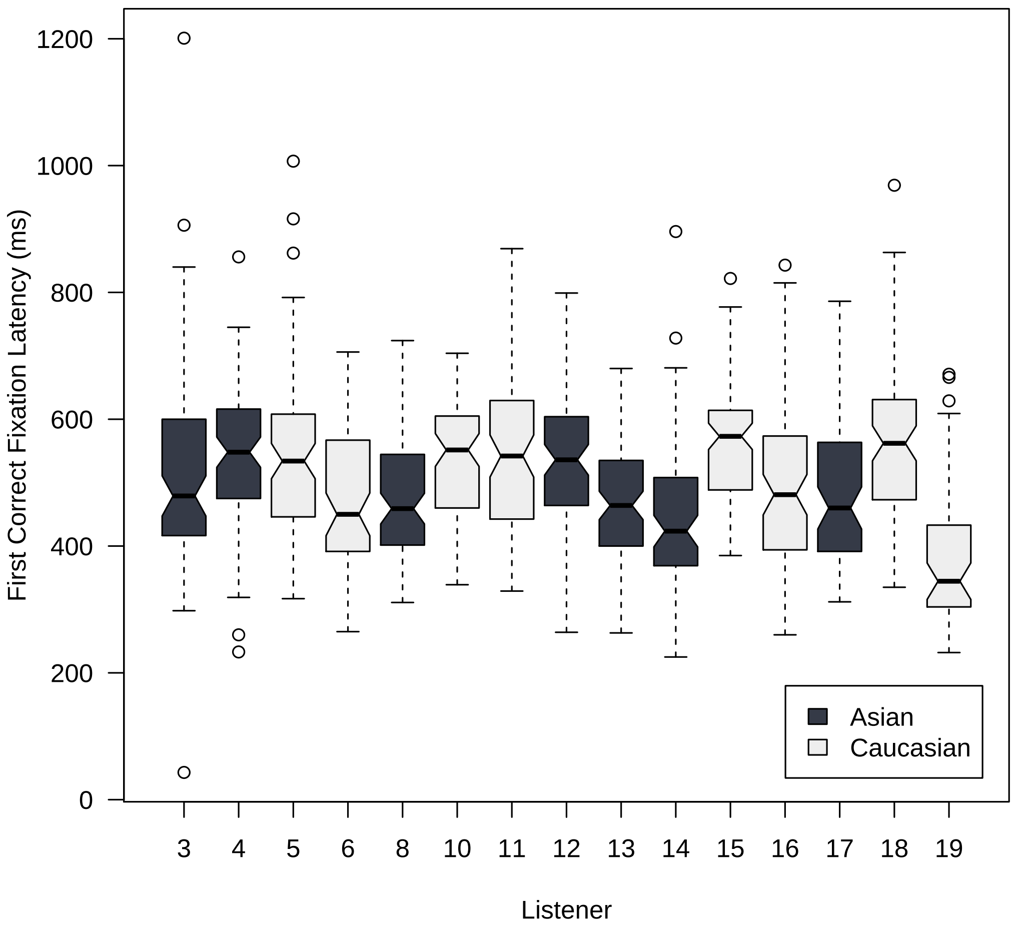

Figure 4.6: **Mean first fixation latencies by listener**

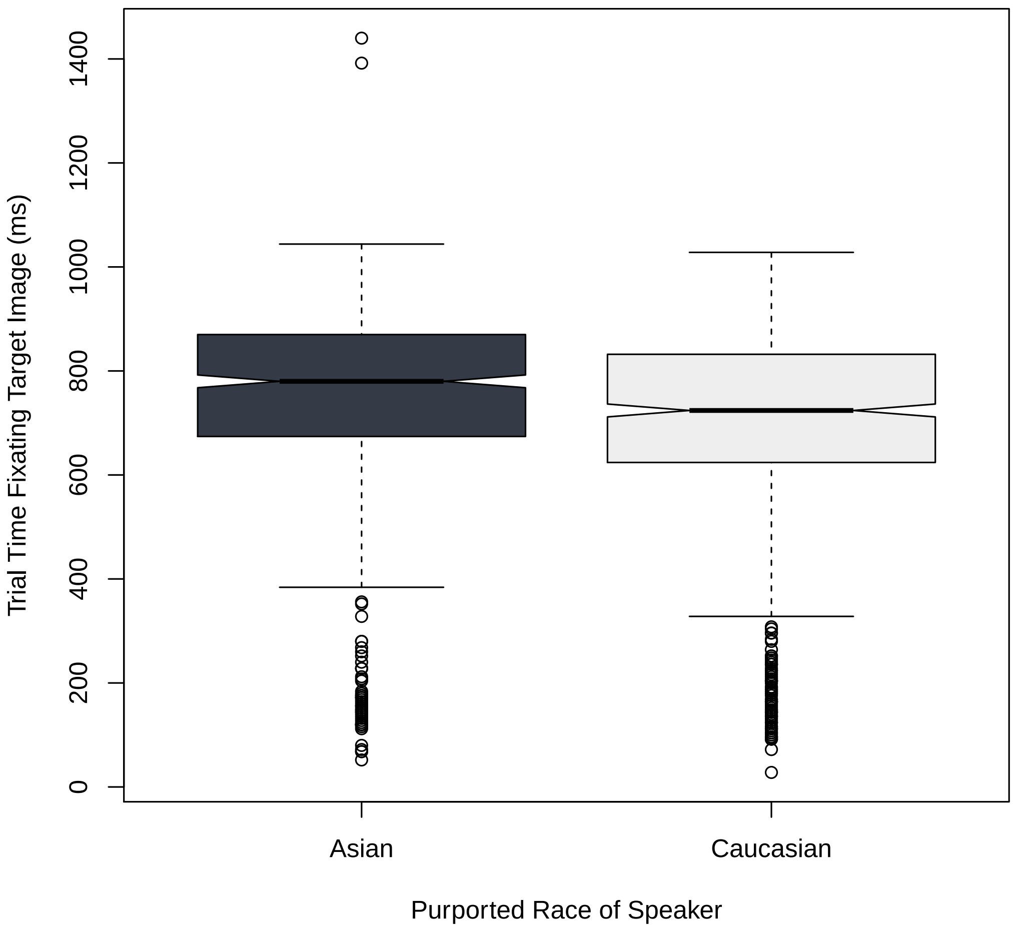

Figure 4.7: **Cumulative time fixating target image by Face condition**

model with, again, Face as a fixed effect and dwell time as the dependent measure. Asian face listeners spent, on average, a greater proportion of the trial time fixating the target image ($\beta = -54.68, p = 0.0219$) with an average of 738ms fixation time versus 682ms for Caucasian face listeners.

One might imagine that the lower mean fixation time in the Caucasian face condition might indicate that these listeners are making more, shorter, fixations overall with potential fixation time therefore lost to saccadic eye movements. This does not appear to the case, though, with Asian condition listeners making, on average, 3.6 fixations over the course of a trial (3.57 to the target image) and Caucasian condition listeners making 3.84 (3.78 of these to the target image). In a linear mixed model with Total Trial Fixations as the dependent variable, Face as a fixed effect, and both Subject and Target as random effects this difference does not achieve significance ($\beta = 0.1738, p = 0.5481$).

This overall significant difference in image dwell times may be attributable entirely to the duration of initial fixations. Across all trials the average first fixation dwell time is 55.5ms longer for Asian condition listeners than for Caucasian condition listeners (735ms vs 679.5ms). This difference is significant ($\beta = -54.42, p = 0.0299$).

4.4 Discussion

Strikingly, at a certain level these results replicate Rubin's findings. Consistent with those findings, listeners in this experiment uniformly reported perceiving a Chinese accent when the face of their purported speaker was Asian. Listeners did not report hearing an accent when the face was Caucasian. It is not clear whether these listeners were really answering the question about the voice they heard or about the face they saw. Listeners in the Asian face condition may well have interpreted the question to mean something like "did you see an Asian face". In any event, the pattern of verbal responses matches those observed by Rubin and suggests that this rather different task may nevertheless be evoking similar types of percepts in listeners.

The remainder of these results are the opposite of the stated predictions. Though not significant, the Asian face condition has first fixation latencies that trend shorter than the Caucasian face condition and is ultimately more accurate (due, I believe, to the longer fixation times) than the Caucasian face condition. This lack of significance in fixation latencies and overall accuracy may be attributable to the low number of participants in this study. It is clearly necessary to extend this experiment to additional participants before any strong claims can be motivated by these results. However, even a null result, the lack of a difference between Asian and Caucasian condition listeners, seems problematic for Rubin and Lippi-Green.

Furthermore, Asian face listeners fixated the target images for a higher proportion of the available trial time and were particularly likely to have longer initial fixations. This seems a clear refutation of Lippi-Green's reduced-attention hypothesis and, as such, reinforces the findings of the transcription task –that listeners in the Asian face condition are as engaged in the task as those in the Caucasian face condition. Merely showing an Asian face to an English-speaking listener does *not* result in diminished attention. The findings reported here do not support Lippi-Green's theory that these listeners are shirking their communicative responsibility or Rubin's interpretation, at least on the basis of speech perception performance, that programs are needed to improve undergraduate attitudes toward non-native speaking instructors.

While it is easy to argue that longer dwell times indicate that the listener is paying attention, it is much more difficult to explain what motivates the difference between Asian and Caucasian groups for identical auditory and visual stimuli. If dwell time is taken to be an indicator that the participant is processing the image being fixated then what would explain the between-group difference in processing? One might also wonder why this additional processing happens *after* the listener has successfully fixated the target image rather than slowing initial fixations as predicted.

It is tempting to hypothesize that this longer fixation time, this longer processing time after lexical access has apparently been achieved, might ultimately lead to the difficulty Rubin's listeners had with the cloze task. If something about the listeners' expectations in the Asian face condition requires them to need more time to process the image and be confident that they have, indeed, selected the correct target there could be a cumulative penalty in a real time listening task that is not visible in a two-item forced choice task like this one. Such a cumulative penalty could possibly explain listeners' difficulty remembering particular words for Rubin's cloze test.

Another observation that can be made from these data is that there does not appear to be a differential use of fine phonetic information. If socioindexical expectation were pre-activating Chinese-labeled exemplars for listeners in the Asian face condition, it seems reasonable to expect that these listeners' first fixation latencies or initial accuracy would suffer as they process the mismatch between acoustic stimulus and their expectations of fine phonetic detail. This lack of an inhibition seems inconsistent with the resonance model of Johnson (2006) in which listeners' Chinese-labeled exemplars are preactivated by socioindexical cues of a Chinese accent. These results are similarly not consistent with a model, like Staum Casasanto (2009a), in which a Bayesian listening strategy leads listeners to shift the prior probabilities over their prelexical and lexical forms in favor of Chinese-accented experience.

However, if these listeners do not have Chinese-labeled exemplars then this result is perfectly in keeping with the predictions of both of the above models. It is necessary to run this experiment again with experienced listeners to examine the time course of the influence of socioindexical expectations when the listener is *experienced* with the variety. The experiment also needs to be expanded to include the same two Face condition pairings but with a Chinese-accented voice. While the results of this particular experiment may seem inconsistent with the resonance and Bayesian bottom-up models, without this additional empirical work I believe it is impossible to rule either of them out.

CHAPTER V

Conclusion

5.1 Summary of Results

The *yes/no* task presented in **Experiment 1** (Chapter II) revealed that experienced and inexperienced listeners are both capable of judging the authenticity of a non-native accent. Experienced listeners were, as one might expect, more capable of identifying authentic Chinese-accented English. Inexperienced listeners, while much more likely to be misled by imitated Chinese than experienced listeners were, still responded 'authentic' to an authentic Chinese accent more often than to any of the other four languages presented. My interpretation was that inexperienced listeners depend more heavily on stereotypical features of what they believe a Chinese accent to sound like than do experienced listeners. When those stereotypical features are performed in an imitated accent or when those features appear in authentic speech, inexperienced listeners can use them.

The speech-in-noise task of **Experiment 2** (Chapter III) addressed whether listeners could use socioindexical expectations to enhance speech perception. Listeners, even inexperienced listeners, presented with an Asian face and a Chinese-accented English voice in a modified matched guise experiment performed a transcription task more accurately than listeners presented with the same voice and either a silhouette or a Caucasian face.

Experienced, Heritage Mandarin Chinese speakers with extensive experience listening to Chinese-accented English were overall more accurate than the inexperienced listeners, but even here a listener presented with an Asian face tended to transcribe more accurately than a listener presented with a Caucasian face and the same authentic Chinese-accented voice and the magnitude of this difference was not significantly different for experienced listeners than for inexperienced listeners (there was no interaction of the Face and Experience conditions in the combined results).

However, within the Experienced listener condition this difference between accuracy in the Asian and Caucasian conditions just missed significance when the sentences being transcribed were highly predictable. This lack of significance may be due to the relatively small number of experienced listeners participating in this study.

Performance given a silhouette designed not to provide indexical hints as to the identity of the speaker patterned differently for inexperienced and experienced listeners when the sentences being transcribed were highly predictable. For inexperienced listeners, the silhouette condition patterned with performance in the Caucasian face and differed significantly from transcription performance in the Asian face condition. Experienced listeners showed a not-significant trend for the opposite pattern: the silhouette condition tended to pattern with performance in the Asian face condition rather than the Caucasian face condition. My interpretation of these results was that listener expectations in the silhouette condition reflect the listeners' default expectations in that population group. This may mean that inexperienced listeners showed a facilitatory effect of Asian face while experienced listeners showed an inhibitory effect of seeing a Caucasian face, though it is not possible to definitively answer this question with the present experiment.

Experiment 3 (Chapter IV) used a visual world eye-tracking paradigm to investigate the time course of the influence of socioindexical expectation on speech perception. What is not clear from performance on the transcription task in Chapter

III is the point (or points) at which socioindexical information exerts an influence on speech perception. It could be, for example, that experienced and inexperienced listeners arrive at similar outcomes on the transcription task but do so by entirely different means –using entirely different cognitive mechanisms or strategies. Furthermore, Lippi-Green (1997) makes a prediction that listeners shirk their portion of the communicative burden by reducing attention to the speech signal. Whether listeners actually show reduced attention is an empirical question and one that eye-tracking is particularly well suited to investigating.

Inexperienced listeners were presented with a Standard American English voice and either an Asian or Caucasian face of the purported speaker. Contrary to predictions and contrary to the usual interpretation of socioindexical perception results in the literature, listeners trended slightly, but not significantly, faster to fixate the correct alternative in a two-alternative forced choice when presented with an Asian face and a Standard American English voice. These listeners also fixated this image for a significantly longer period. Strikingly, listeners in the Asian face condition unanimously reported hearing an accent while listeners in the Caucasian condition unanimously did not.

These results were interpreted to refute Lippi-Green (1997) and Rubin (1992)'s claim that listeners presented with an Asian face pay less attention to the speech of their interlocutor. Furthermore, these results are not in keeping with the prediction that social expectations consistent with a speech stimulus will facilitate perception of that stimulus by pre-activating stored exemplars.

5.1.1 The Exemplar Hypothesis

This dissertation has presented a series of experiments designed to investigate listeners' ability to use socioindexical expectations during speech perception. I have argued that attributing observed socioindexical perception effects to stored episodic traces is not always motivated by the experiments used to find them. Often, these attributions are due to an inference that an observable influence of socioindexical information on perception performance must necessarily be attributable to stored experience. I believe I have shown that this is not necessarily the case and that listeners lacking experience with Chinese-accented English can also show strong influences of socioindexical expectations –even using these expectations to enhance performance on a speech-in-noise task.

Ultimately this dissertation is not, and does not aim to be, a refutation of exemplar models. What I have sought to demonstrate is merely that experimental evidence of socioindexical influence on speech perception is not necessarily evidence that listeners store detailed episodic traces of perceptual experience. These results constrain the set of possible cognitive mechanisms underlying perception to those that afford experiential learning and the influence of social knowledge on the processing of sensory information. However, there is a vast gulf of missing experimental evidence between these constraints and a connection to stored episodic traces. This dissertation has attempted to cross that gulf by testing the predictions of the particularly well-elaborated model in Johnson (2006), and the evidence does not support the predictions of the model. Social knowledge does not appear to shape bottom-up speech perception by pre-activating those exemplars associated with an activated social category. The fixation patterns observed in, for example, Chapter IV could easily be associated with a model in which bottom-up processing of the acoustic signal is mediated by a socially-informed top-down pruning after lexical entries consistent with the acoustic input are activated.

In the future I will build on this research to investigate the assumption that classification judgments will be able to draw directly on fine phonetic detail even when made several moments after listeners process a target stimulus item. It is assumed in, for example, Niedzielski (1999) that listeners' early judgments of indexical phonetic detail will remain constant throughout a listening task and that these judgments will not be shifted, supplanted or, indeed, created by later, higher level cognitive processes associated with that task. The results of the eye-tracking experiment at least cast doubt on this assumption for accented speech and I hope that further work investigating socioindexical results with more online tasks will prove revealing.

5.1.2 The Negative Bias Hypothesis

Finally, I have looked particularly closely at the results of Rubin (1992). Rubin, and later Lippi-Green (1997), interpreted the results of a Matched Guise Test involving Asian and Caucasian faces as evidence that listener bias, particularly racial bias, can lead to reduced attention and reduced engagement in the role of listener. I have shown that this interpretation can not be correct. Listeners' performance can, in fact, be *improved* by the presentation of an Asian face when the speech is Chinese-accented. Improved performance is not consistent with Rubin and Lippi-Green's interpretations. Moreover, the eye-tracking data, a measure of visual attention, show no evidence of reduced attention when listeners have been primed with an Asian face.

5.2 Conclusion

Speech perception research is by its very nature an interdisciplinary enterprise drawing on expertise from linguistics, psychology, speech and hearing and computational linguistics. A field that began as largely the domain of telephone engineers (Licklider and Miller, 1951) has evolved again and again to embrace new techniques, models and technologies. Over the past decade, the field has evolved again to incorporate an investigation of the variation that has traditionally been the focus of sociolinguists. This dissertation has points of contact with each of these fields and with linguistic anthropology.

The questions asked here and my attempts to answer them reflect, I hope, my training as a phonetician in a linguistics department. What listeners know when they know a language, how that knowledge is stored in the mind and particularly how speaker/listeners use that knowledge when producing and perceiving speech are the questions that drive my research. This dissertation adds more evidence to the position that listeners, both experienced and otherwise, have socioindexical knowledge that drives expectations during speech perception. This evidence is inconsistent with linguistic theories that attempt to divorce linguistic competence from social factors.

In the introduction and in my experimental designs I have taken the position that the fundamental questions of socioindexical speech perception represent a natural continuation of decades of research into how listeners robustly map highly variable acoustic information onto mental representations. Rather than being a new field of investigation, socioindexical speech perception is more correctly understood as bringing a new appreciation for identity and social relationships to speech perception.

APPENDICES

APPENDIX A

Experiment 2: Inexperienced Transcription Keyword Errors: High Predictability Condition

animals

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	m	1	and	1	annuals
		1	annuals	1	monster
				1	sale

bird

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	a	1	a	1	canister
1	j	1	at	1	cant
		1	canister	1	cat
		1	for	1	infered
		1	kjl	1	picture
				1	words

bomb

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	bunk	2	ball	1	ball
		1	jumped	1	it
		1	travbolak		

cents

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	fifths	3	steps	2	steps
1	steps	2	25	1	6
1	teps	1	NA	1	biceps
		1	degrees	1	fifths
		1	first	1	quarter
		1	her	1	them
		1	quater	1	what
		1	sips		
		1	tricks		

coach

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	college	4	college	4	college
1	couch			2	couch
1	culture				

coffee

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	Many		

days

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	NA	1	ago	2	NA
1	Feburary	1	bad	2	talk
1	bangs	1	childs	2	today
1	burberry	1	favorite	1	?
1	extra	1	february	1	april
1	jek	1	febuary	1	at
1	space	1	hdhryhf	1	february
1	things	1	huh	1	nothing
1	top	1	steak	1	tall
1	twelve	1	tempory	1	told
		1	three		
		1	to		
		1	today		
		1	top		

dinner

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	pudding	1	cuardinos	1	big
		1	had	1	home
				1	something

family

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	parents	1	parents	1	offended
				1	talking

fast

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	go	4	go	4	go
3	where	2	where	4	where
1	helps	1	else	1	NA
1	please	1	jdjdjd	1	at
		1	raced	1	does
		1	track	1	the
		1	with	1	work

father

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	brother	3	brother	6	mother
1	NA	2	engine	3	brother
1	boy	2	helped	1	boys
1	bus	1	boys	1	bus
1	daughter	1	excercise	1	interesting
1	en	1	exercise	1	something
1	enterprise	1	himself		
1	wrist	1	mother		
		1	smarter		
		1	spot		
		1	they		

feet

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	feel		
		1	you		

grass

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	NA	4	she	2	NA
2	she	1	bread	2	eat
1	bread	1	breath	1	bread
1	eats	1	feel	1	hassel
1	fresh	1	grabs	1	rained
1	grace	1	ran	1	she
1	graph	1	sdf;jk		
1	the	1	was		
		1	wept		

head

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	NA		

juice

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	trees	2	trees	2	trees
		1	NA		
		1	jump		

leaves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	of	2	NA	1	plan
1	operate	1	frisbees	1	plant
		1	fully	1	seeds
		1	grade	1	to
		1	windy		

necks

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	next	1	accent	1	neck
1	wrong	1	dfkvx	1	next

pie

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	had	2	ever	4	had
1	NA	2	everyone	3	ever
1	aderpon	2	had	1	?
1	apricot	1	advertise	1	NA
1	does	1	advertised	1	apples
1	eva	1	everybody	1	apricot
1	everyone	1	everything	1	eat
1	everything	1	has	1	good
1	qua	1	tried		
1	saw	1	tron		
1	tried				

sky

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	class	1	NA
				1	clouds
				1	ska
				1	skies

sleeves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	NA	4	leaves	3	NA
2	leaves	2	NA	2	leaves
1	clips	2	was	2	plates
1	grapes	1	?	2	played
1	leaf	1	a	1	clips
1	lips	1	bought	1	legs
1	nkjwe	1	clean	1	lips
1	posture	1	flakes	1	lives
1	should	1	hgber	1	place
1	sleep	1	late	1	shirt
1	sponsee	1	slaves	1	sport
		1	your		

sport

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
14	spot	14	spot	11	spot
		1	NA	1	adventurous
		1	iwurfr	1	bot
		1	plot	1	bought
		1	response	1	dangerous
				1	tot

story

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	NA	1	dary	1	dari
1	diary	1	diary	1	diary
1	dolly	1	sorry	1	stories
		1	talladari	1	tells
		1	tells		

time

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	tide	1	watch
		1	watch		

trees

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	herro	2	NA
		1	natural	1	interest
		1	oiefviu	1	interests
		1	tree	1	the
		1	truths	1	tress
		1	vesbute		

water

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				1	what

wrist

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	watch	1	ball	2	watch
		1	brace	1	watched
		1	movies	1	white
		1	outerspace		
		1	roll		
		1	tuesday		
		1	watched		

yellow

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	lemon				
1	limo				

APPENDIX B

Experiment 2: Inexperienced Transcription Keyword Errors: Low Predictability Condition

animals

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	NA	1	analyst	1	annual
1	analyst	1	angles	1	it
1	that	1	help	1	once
		1	jfj	1	session
		1	section	1	the
		1	well	1	there

bird

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	agirl	1	early	1	NA
1	apoltuabur	1	too	1	girl
1	girl				

bomb

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	about	3	bone	5	phone
2	phone	3	bones	2	NA
1	bo	1	about	2	about
1	bone	1	ball	1	bond
1	bones	1	balls	1	bone
1	bonk	1	bunk	1	bowl
1	book	1	phone	1	talked
1	boy	1	phones	1	taught
1	ecology	1	revolt	1	the
1	orange	1	the		
1	phone/bone				
1	the				

cents

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	NA	3	fence	6	fence
2	fence	2	the	2	NA
2	punched	1	NA	1	defense
1	at	1	assest	1	desk
1	defense	1	defence	1	pointed
1	defense/elephants	1	defense	1	resets
1	distance	1	desk	1	sentence
1	k	1	fists	1	the
1	ketz	1	hands	1	tickets
1	set	1	heads		
1	states	1	kdds		
1	the	1	next		
		1	once		
		1	since		
		1	weekends		

clock

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	prop			1	plant

coach

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
6	couch	11	college	8	couch
5	college	5	couch	6	college
1	collegee	1	colors		
1	courage				
1	culture				

coffee

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	copy	1	ducati	2	NA
1	j			1	about
				1	parking
				1	the

days

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	NA	1	bags	1	base
1	age	1	bays	1	here
1	babies	1	dates	1	they
1	bake	1	here	1	year
		1	many		
		1	vague		

dinner

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	enough			1	about
				1	the

family

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	NA	3	vanity	2	NA
2	about	1	NA	2	about
1	Bambi	1	about	2	damage
1	bambi	1	bambi	1	age
1	dam	1	dancer	1	contaminate
		1	examining	1	damage?
		1	teacher	1	teacher
				1	vanity
				1	vanny

fast

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	helps	1	rust	2	frost
				1	fun
				1	is

father

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	brother	2	brother	2	brother
		2	helped	2	farther
		1	Mom	1	NA
		1	harder	1	mother
		1	huh	1	the
		1	platter		

feet

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	picture	5	pictures	4	her
2	carpet	2	picture	3	picture
2	wrist	1	NA	2	face
1	bit	1	a	1	NA
1	her	1	faace	1	at
1	kids	1	figure	1	dad
1	of	1	her	1	fit
1	perfect	1	hubet	1	storyl
1	pictures	1	outfit	1	watch
1	temperature				
1	turbin				

grass

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	graph	5	graph	5	graph
2	ground	3	ground	2	ground
1	breath	1	ref	1	brat
1	clock			1	breath
1	giraffe			1	that
1	grad			1	undergrad
1	the				

head

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	hat	1	hat	4	hat
1	j	1	jsdgds		

juice

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	jewels	1	jewls	2	you
1	something	1	shoes	1	NA
1	trees	1	trees	1	as
		1	troops	1	shoes
		1	truth	1	troops

leaves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	me	3	the	4	me
1	cat	1	about	2	the
1	catobodita	1	leeks	1	leaving
1	league	1	m	1	neat
1	lip	1	me	1	these
1	the			1	we
				1	weed

necks

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	next	1	about	2	NA
3	about	1	accent	2	about
1	accent	1	ask	2	xanex
1	accents	1	bandaids	1	Vanax
1	aids	1	ex	1	ace
1	annex	1	fedex	1	bannex
1	ben	1	five	1	bed
1	nests	1	legs	1	detective
1	x	1	neice	1	have
		1	nests		
		1	next		
		1	them		
		1	touched		
		1	x		
		1	xanax		
		1	xanax?		

pie

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	hike	2	about	1	I
1	why?	2	hide	1	NA
		1	behind	1	pine
		1	dad	1	white
		1	her		

sheets

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	trees	4	trees	5	trees
1	about	2	about	1	about
1	ballstee	1	chicks	1	ball
1	cheese	1	shapes	1	chicks
1	dad	1	she	1	dad
1	shirts	1	sheep	1	her
1	shit	1	shit	1	sheep
1	shoes	1	shoes	1	shoes
1	street	1	streets	1	street
		1	things	1	streets
				1	things

sky

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
6	die	7	guy	5	guy
5	guy	3	die	2	NA
1	NA	1	NA	2	die
1	gu	1	high	1	dad
1	guide	1	sty	1	dive

sleeves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
6	leaves	5	leaves	5	leaves
2	lips	3	names	4	lips
2	the	1	NA	3	legs
1	flakes	1	flames	2	the
1	flips	1	her	1	lake
1	league	1	lips	1	player
		1	look	1	something
		1	looked		
		1	plays		
		1	space		
		1	the		
		1	way		

sport

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
13	spot	17	spot	17	spot
1	NA	1	a		
1	spoty				

story

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	star	3	stars	5	stars
1	dad	1	NA	3	diary
1	dawy	1	asljhfd	1	dad
1	diary	1	at	1	darling
1	dolly	1	dadis	1	daughter
1	the	1	darling	1	star
		1	diary	1	the
		1	is		
		1	starry		
		1	the		

time

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	type	7	type	1	kind
1	NA	1	tie	1	type

trees

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	truce	3	truth	2	NA
2	truth	1	points	2	tricks
2	truths	1	tricks	1	hear
1	NA	1	truths	1	three
1	Trees			1	triste
1	Truth			1	truth
1	said				
1	treats				

water

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	fresh				

week

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	wig	3	twig	4	wig
2	drink	2	drink	1	drink
1	trick	2	wig		
		1	treat		

wrist

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	the	2	with	2	at
2	witch	1	NA	2	him
1	NA	1	add	2	picture
1	at	1	at	2	the
1	kjdfjkher	1	dad	2	with
1	picture	1	her	1	NA
1	watch	1	jfds	1	add
1	which	1	look	1	is
1	wist	1	me	1	which
1	with	1	picture		
		1	the		
		1	waist		
		1	which		
		1	whisk		
		1	wisp		
		1	witch		

yellow

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	that	1	exists	1	NA
1	cat	1	here	1	dad
1	ill	1	idiot	1	is
1	yuck	1	old	1	picutre
		1	sear	1	that
				1	think

APPENDIX C

Experiment 2: Experienced Transcription Keyword Errors: High Predictability Condition

animals

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	value	1	annual

bird

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	bored			1	peach
1	canned				
1	teacher				
1	word				

bomb

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	bunk	1	abound	1	ball
1	bonk	1	ball	1	bunk
1	played	1	bond	1	place
				1	played

cents

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	five	1	steps	1	steps
1	sets				
1	steps				

coach

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	college			1	college
				1	couch

days

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	12th	1	NA	1	arces
1	blank	1	thirty	1	change
1	thirty			1	today

dinner

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	pudding		

family

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	friendly				

fast

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	where	2	go	2	go
2	go	2	where	1	asdf
1	anywhere	1	in	1	i
1	kkk			1	press
				1	where
				1	wherever
				1	with

father

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	bus	3	brother	2	bus
3	mother	1	bus	1	and
2	brother	1	mother	1	boy
1	advice			1	brother
1	rice			1	inside
1	son			1	mother
1	sons				

feet

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				1	fret

grass

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	bread	1	grabs	1	bread
1	grasped	1	grasps	1	breath
		1	she	1	feels
				1	graphs
				1	it

juice

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	trees	1	juicy

leaves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	of		

necks

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	nest	1	neck	1	neck
		1	nest		
		1	of		

pie

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	found			1	everything
1	had			1	prime
1	jjj			1	time
1	pot			1	tried
1	required			1	try
1	whined				

sleeves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	leaves	1	NA	2	leaves
1	Sports	1	shirt	1	boy
1	blank			1	has
1	lease			1	leagues
1	nitrogen				
1	responder				

sport

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
7	spot	5	spot	6	spot
1	bot	1	boat	1	bot

story

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	stories	1	stories	1	stories

time

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	tide	1	times
				1	type

trees

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	intringue	1	tree	1	intrigued
				1	intrigues
				1	natural

water

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				2	soap

wrist

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	sea	1	face	2	history
1	what	1	on	1	face
1	wirst	1	street	1	ice
		1	watch		

yellow

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				1	seattle

APPENDIX D

Experiment 2: Experienced Transcription Keyword Errors: Low Predictability Condition

animals

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	is	1	annually	2	angle
		1	of	1	said
		1	the		
		1	wanted		

bird

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		2	birds	1	birds
		1	NA	1	the

bomb

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
8	phone	7	phone	3	phone
2	bone	1	boe	1	about
1	bones			1	boat
1	forms			1	bones
				1	book
				1	performance
				1	uniforms

cents

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	desk	1	NA	1	distance
2	sets	1	fence	1	guest
2	steps	1	sacks	1	know
1	NA	1	sense	1	new
1	defense	1	sets	1	sense
1	insects	1	test	1	sets
1	looset?	1	the	1	yourself
1	sentence				
1	set				

clock

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	blank			1	car

coach

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
6	college	5	couch	5	college
4	couch	1	college	4	couch

coffee

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				1	copy

days

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	dates	1	babies		
1	ways	1	dates		

dinner

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
				1	dinnner

family

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	fantasy	1	famine	1	fantasy
1	vanity	1	the	1	that
				1	the

fast

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	fine				

father

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	feather	1	brother	2	farther
1	further				
1	mom				

feet

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	picture	1	a	1	her
1	a	1	face	1	outfit
1	carpet	1	picture	1	picture
1	fridge	1	tuppet	1	story
1	outfit	1	watch		
1	tuffet	1	wich		

grass

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	graph	2	graph	1	asdf
1	giraffe			1	autograph
1	grad			1	garage
				1	ground
				1	the

head

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
		1	hat	3	hat

juice

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	jews	1	jewels	1	jews
1	juie			1	truths

leaves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	the	1	about	1	liv
1	believe	1	me	1	the
		1	the	1	thieves

necks

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
4	next	3	next	1	accident
1	Ben	2	about	1	band
1	about	1	ex	1	excessively
1	x	1	nest	1	next
				1	the

pie

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	hi				
1	mom				

sheets

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	sheep	3	about	3	about
3	street	1	shit	1	chic
1	she	1	trees	1	she
1	sheeps			1	sheep
				1	trees

sky

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	guy	1	bought/about	3	guy
2	died	1	decorated	2	die
1	read	1	guide		
		1	guy		

sleeves

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
3	leaves	1	leaves	2	leaves
1	knees	1	place	2	slaves
1	names	1	plate	1	as
1	slave	1	the	1	label
1	sleep			1	the
1	sleeve				

sport

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
12	spot	8	spot	7	spot
				1	spote

story

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
2	stars	2	stories	4	stories
1	look	1	sorry	1	April
1	starry	1	starry	1	starry
		1	the	1	the

time

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	type	1	type	2	tie
				1	type

trees

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	juice	1		1	truce
1	treats	1	trip	1	tuse
1	truths				

week

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	drink	1	drink	1	favorite
1	twick	1	trick	1	wig
1	wig	1	twig		

wrist

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	catch	2	with	2	which
1	picture	1	the	1	add
1	that	1	wheat	1	the
1	the			1	with
1	waist				
1	we				
1	which				

yellow

Asian Face		Silhouette		Caucasian Face	
count	response	count	response	count	response
1	a	1	here	1	Seattle
1	end	1	its	1	egg
1	here	1	that	1	is
1	other			1	zero
1	yelp				
1	younger				

APPENDIX E

Experiment 3: Visual Stimuli for Eye Tracking Experiments

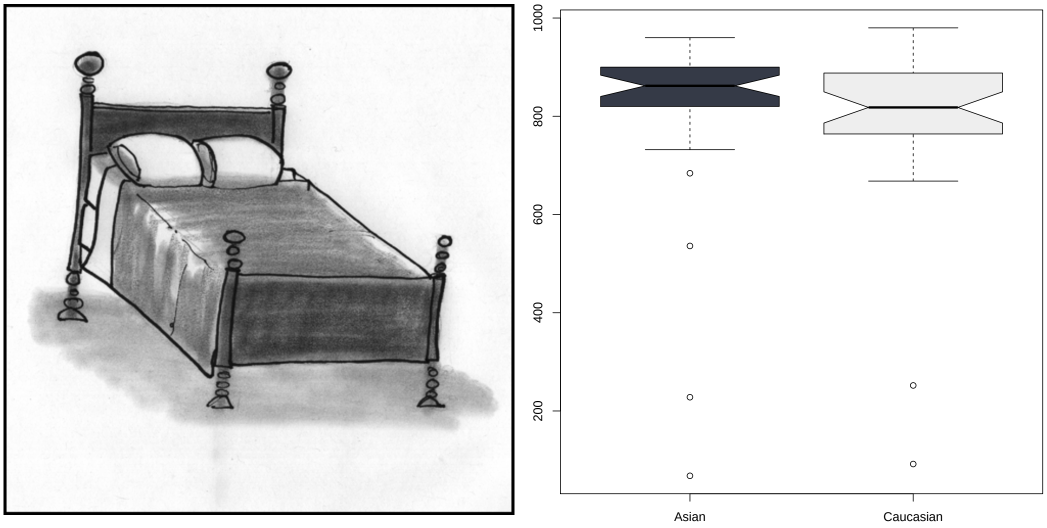

Figure E.1: Target image for *bed* and mean cumulative dwell time by Face.

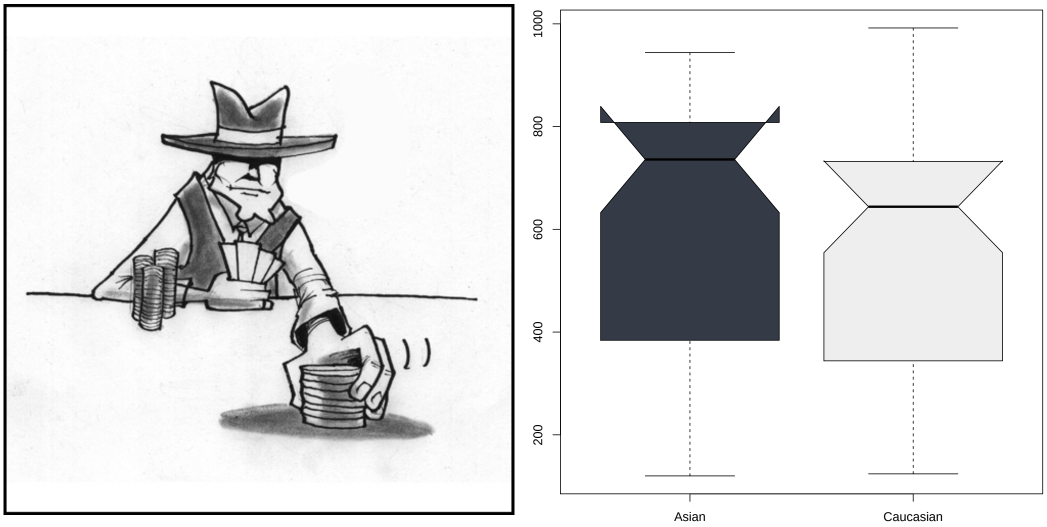

Figure E.2: Target image for *bet* and mean cumulative dwell time by Face.

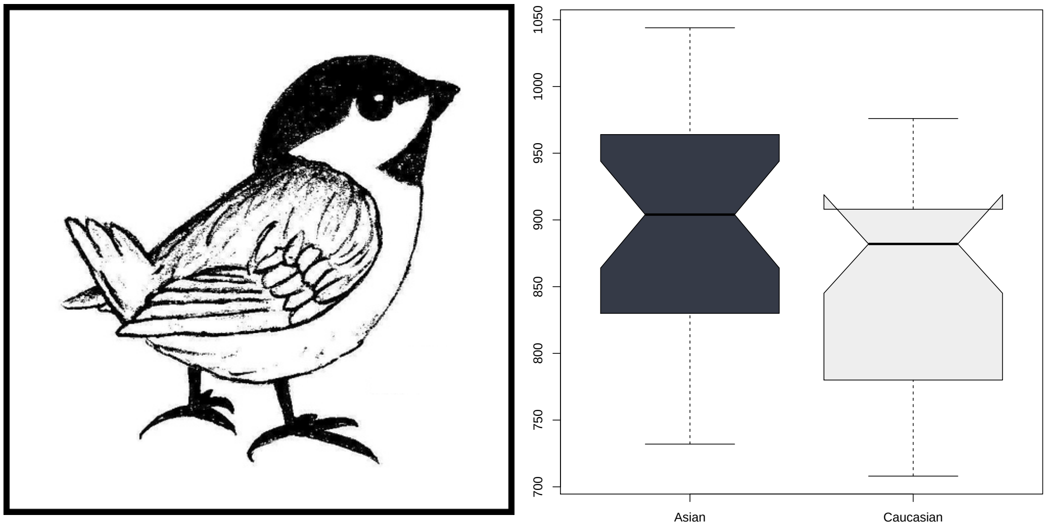

Figure E.3: Target image for *bird* and mean cumulative dwell time by Face.

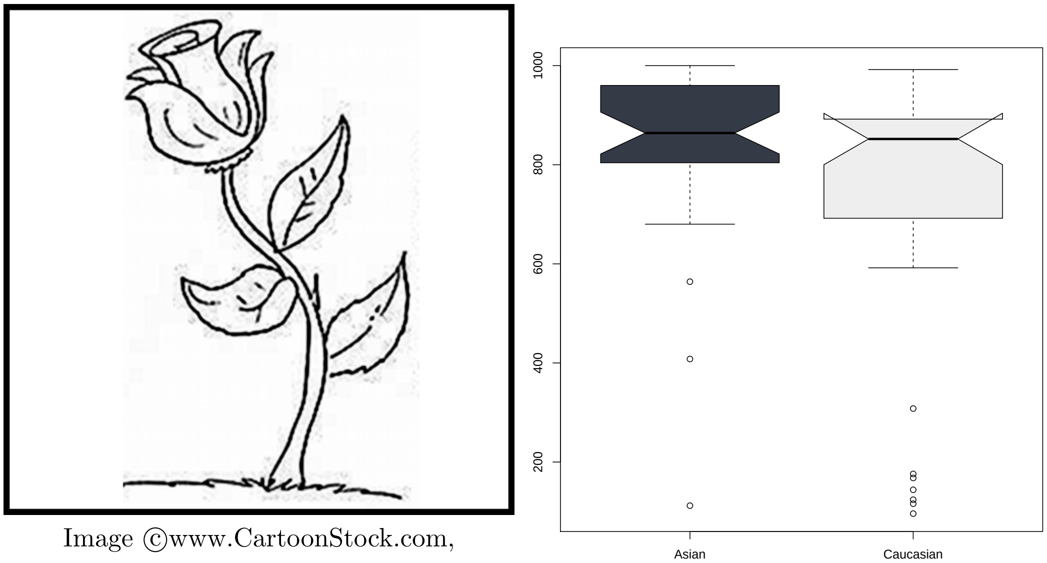

Image ©www.CartoonStock.com,
used with permission.

Figure E.4: Target image for *bud* and mean cumulative dwell time by Face.

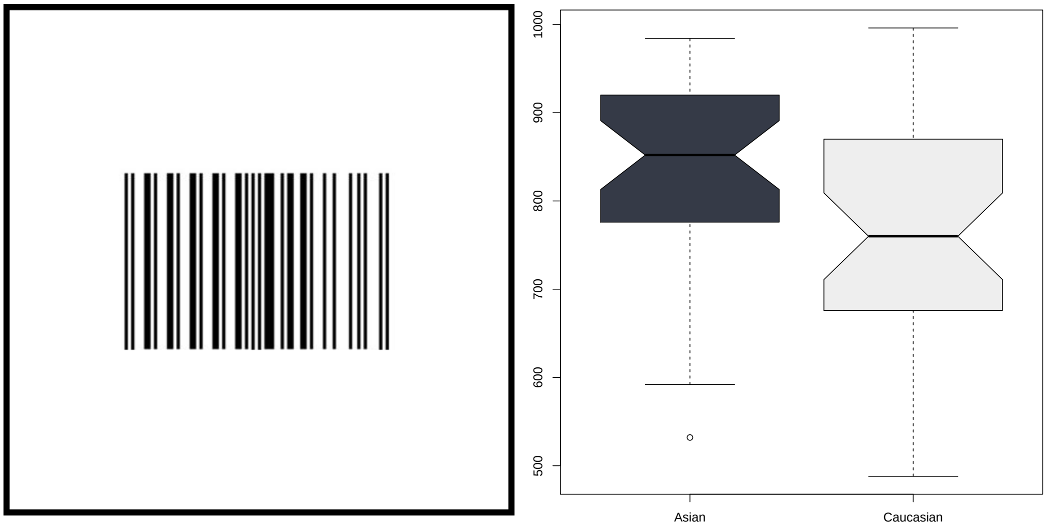

Figure E.5: Target image for *code* and mean cumulative dwell time by Face.

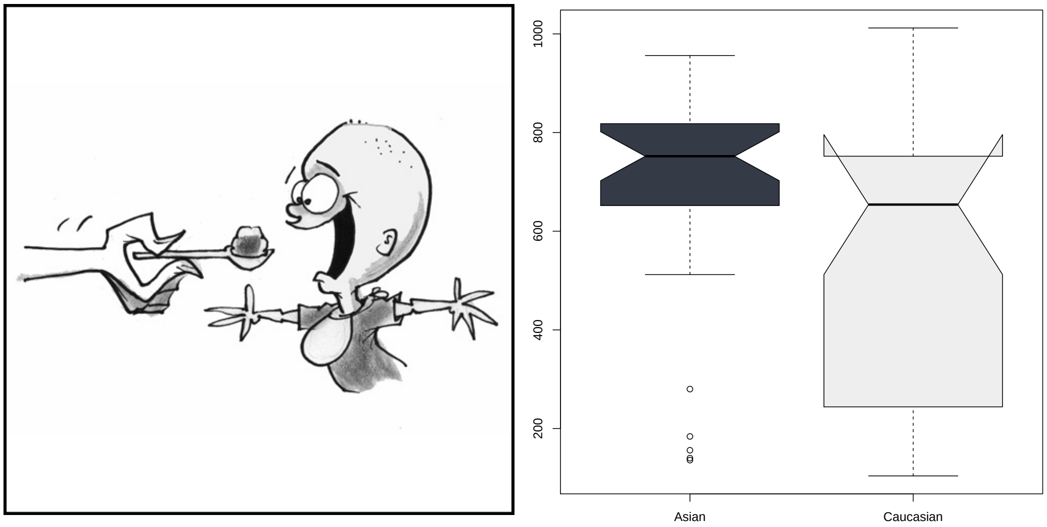

Figure E.6: Target image for *feed* and mean cumulative dwell time by Face.

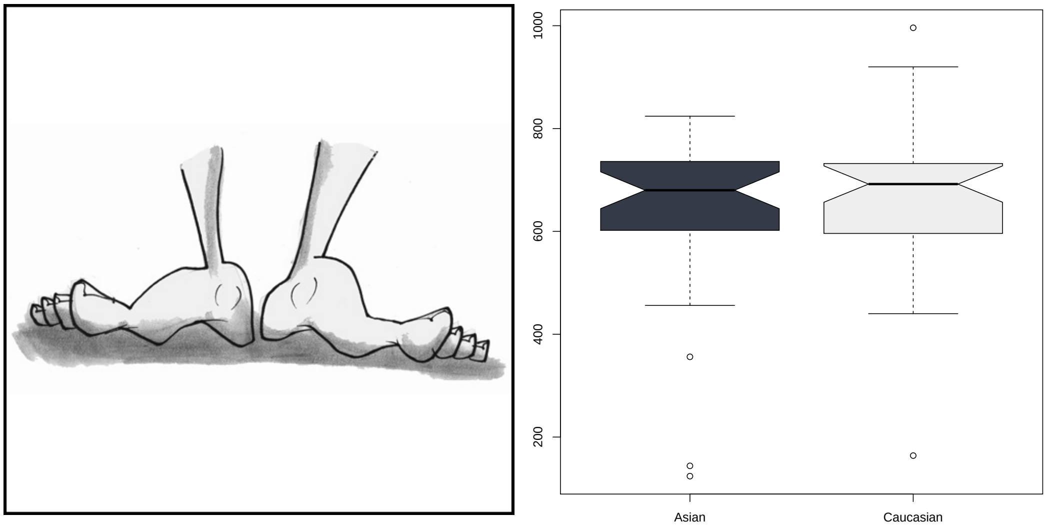

Figure E.7: Target image for *feet* and mean cumulative dwell time by Face.

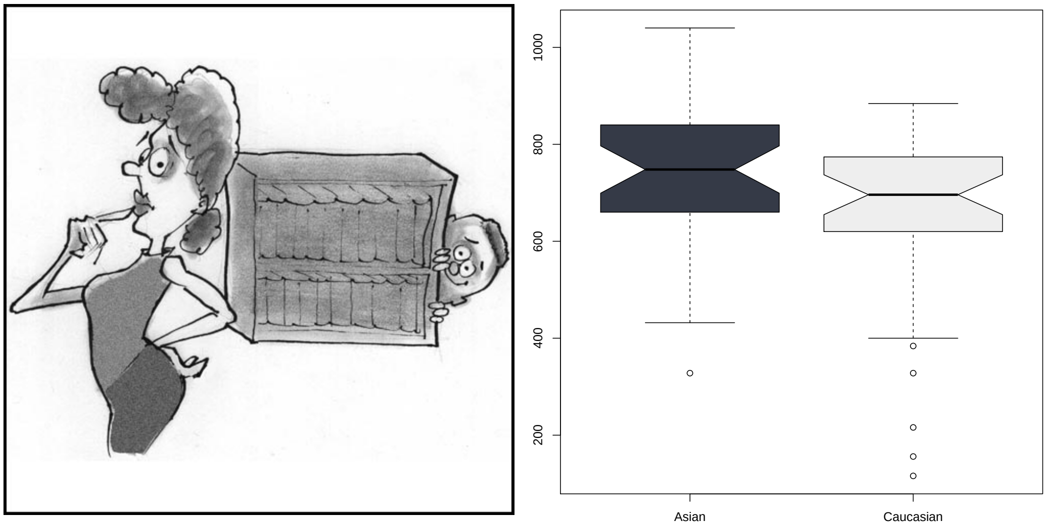

Figure E.8: Target image for *hid* and mean cumulative dwell time by Face.

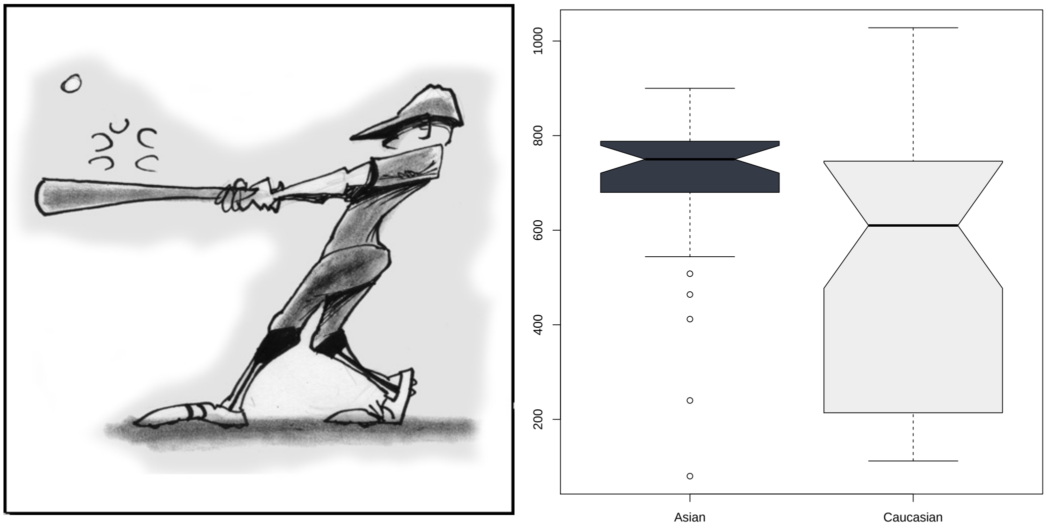

Figure E.9: Target image for *hit* and mean cumulative dwell time by Face.

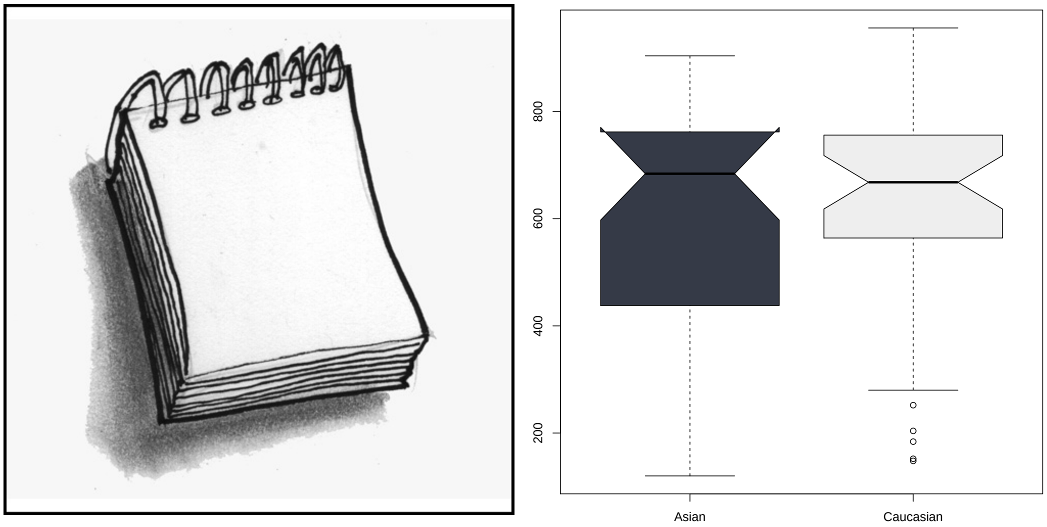

Figure E.10: Target image for *pad* and mean cumulative dwell time by Face.

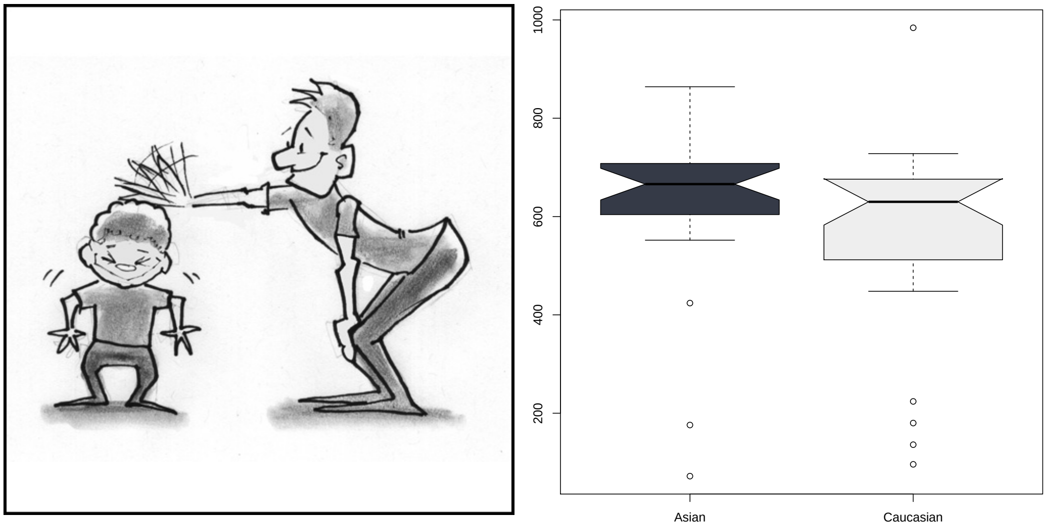

Figure E.11: Target image for *pat* and mean cumulative dwell time by Face.

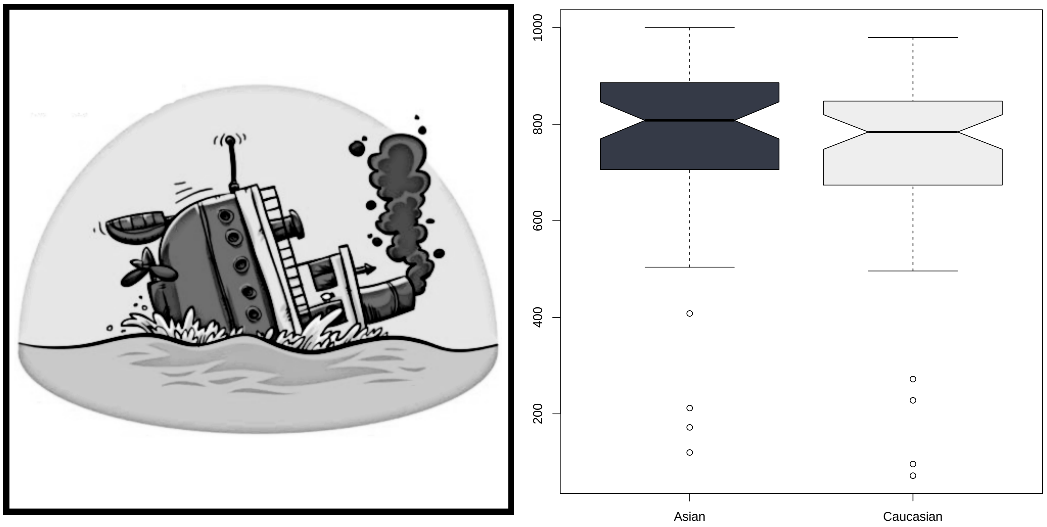

Figure E.12: Target image for *sank* and mean cumulative dwell time by Face.

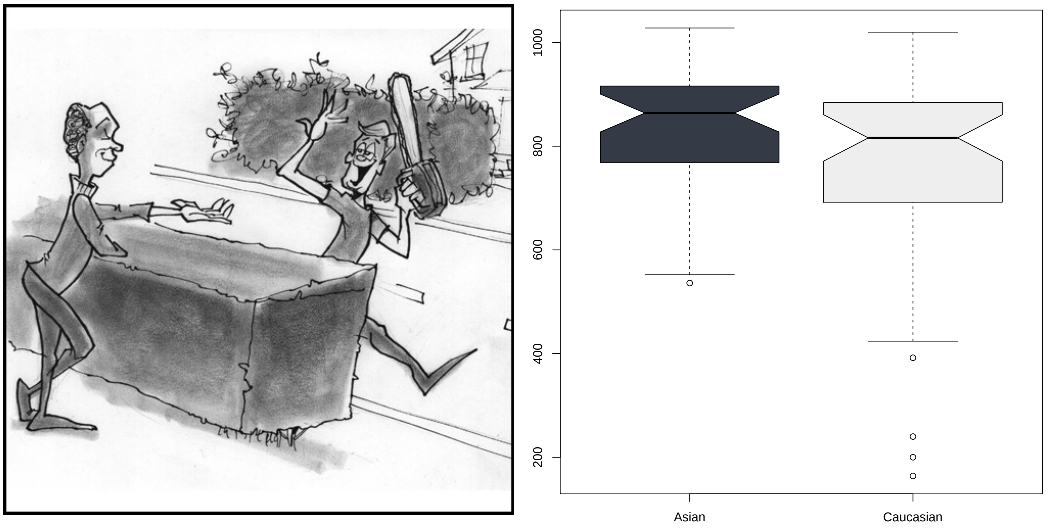

Figure E.13: Target image for *thank* and mean cumulative dwell time by Face.

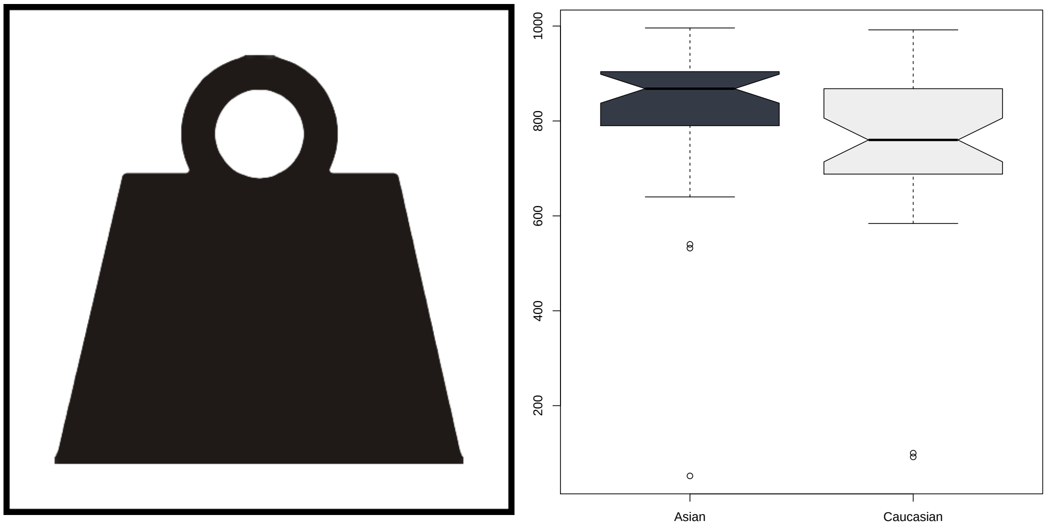

Figure E.14: Target image for *ton* and mean cumulative dwell time by Face.

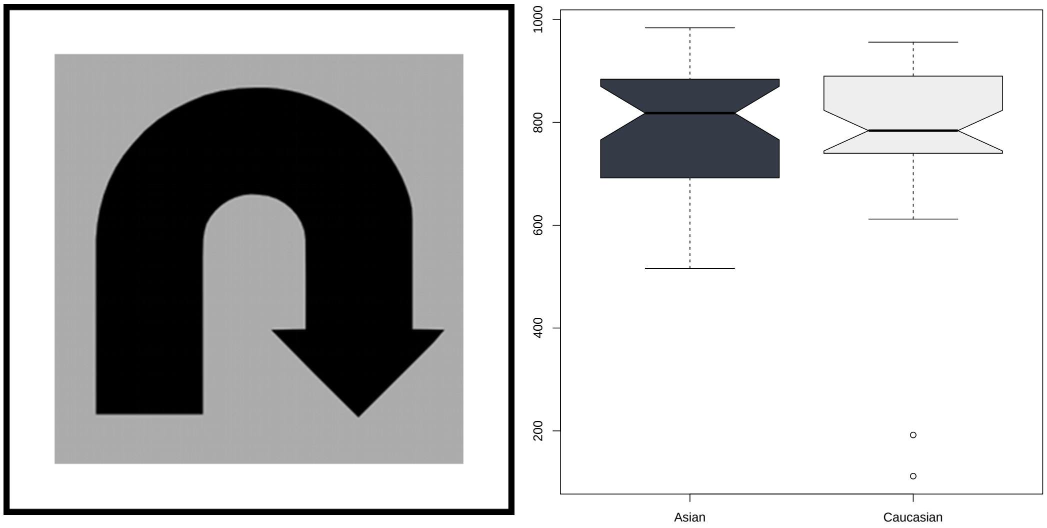

Figure E.15: Target image for *turn* and mean cumulative dwell time by Face.

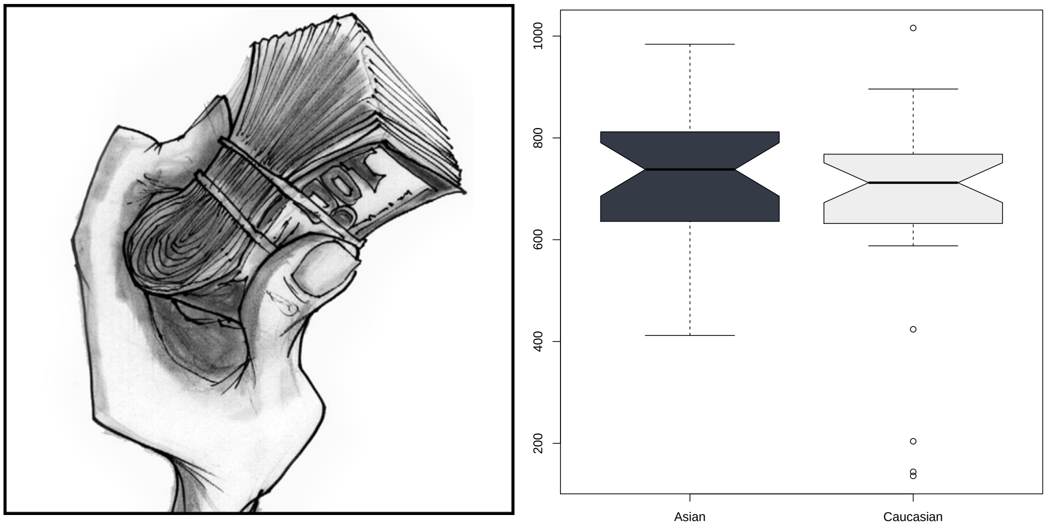

Figure E.16: Target image for *wad* and mean cumulative dwell time by Face.

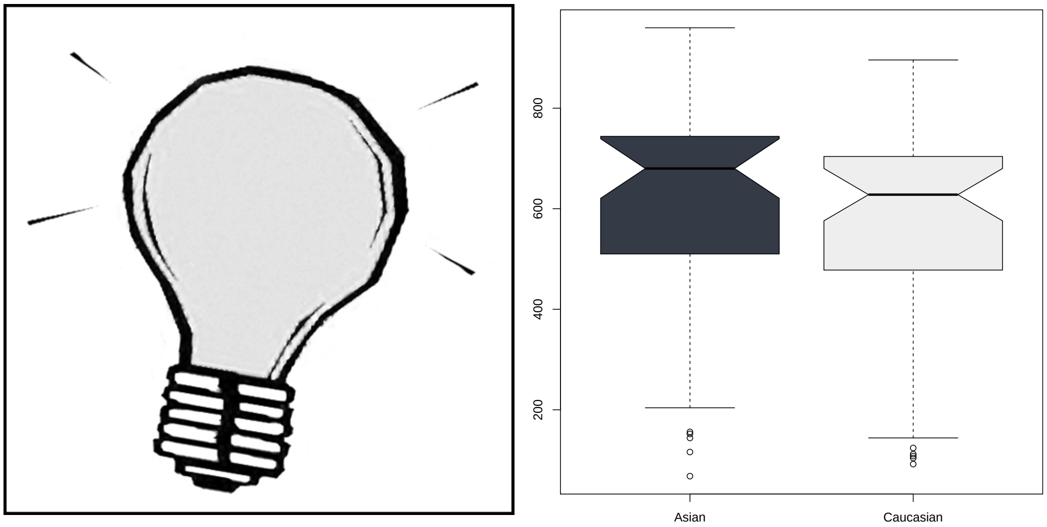

Figure E.17: Target image for *watt* and mean cumulative dwell time by Face.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abramson, A. and Lisker, L. Discriminability along the voicing continuum: cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences Prague 1967*, pages 569–73. Academia, Prague, 1970.

Adank, P., Smits, R. and van Hout, R. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5):3099–3107, 2004.

Allopenna, P. D., Magnuson, J. and Tanenhaus, M. Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language*, 38:419–439, 1998.

Appelbaum, I. The dogma of isomorphism: A case study from speech perception. *Philosophy of Science*, 66:pp. S250–S259, 1999. ISSN 00318248.

Au, T., Knightly, L., Jun, S. and Oh, J. Overhearing a language during childhood. *Psychological Science*, 13(3):238–43, 2002.

Baayen, R. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge University Press, Cambridge, 2008.

Baayen, R. H., Piepenbrock, R. and Rijn, H. V. The celex lexical data base on cd-rom. 1993.

Babel, A. M. *Contact and Contrast in Valley Spanish.* Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2010.

Babel, M. *Phonetic and social selectivity in speech accommodation.* Ph.D. thesis, University of California, Berkeley, Berkeley, CA, 2009.

Babel, M. and Johnson, K. Accessing psycho-acoustic perception and language-specific perception with speech sounds. *Laboratory Phonology*, 1(1):179–205, 2010. doi:10.1515/LABPHON.2010.009.

Balota, D., Yap, M., Hutchison, K., Cortese, M., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G. and Treiman, R. The english lexicon project. *Behavior Research Methods*, 39(3):445–459, 2007. doi:10.3758/BF03193014.

Bates, D., Maechler, M. and Bolker, B. *lme4: Linear mixed-effects models using S4 classes*, 2011. R package version 0.999375-39.

Beddor, P. S., Boland, J., Coetzee, A. and McGowan, K. B. The perceptual time course of coarticulatory nasalization (poster). In *158th Meeting of the Acoustical Society of America*. 2009.

Best, C. A direct realist perspective on cross-language speech perception. In Strange, W. and Jenkins, J. J., editors, *Cross-language speech perception*, pages 171–204. York Press, Timonium, MD, 1995.

Bloomfield, L. *Language*. Holt, New York, 1933.

Bradlow, A. and Alexander, J. Semantic-contextual and acoustic-phonetic enhancements for english sentence-in-noise recognition by native and non-native listeners. *Journal of the Acoustical Society of America*, 121(4):2339–2349, 2007.

Bruce, D. J. The effect of listeners' anticipations on the intelligibility of heard speech. *Language and Speech*, 1(2):79–97, 1958.

Bucholtz, M. and Hall, K. Locating identity in language. In Llamas, C. and Watt, D., editors, *Language and Identities*, pages 18–28. Edinburgh University Press, 2010.

Campbell-Kibbler, K. *Listener perceptions of sociolinguistic variables: The case of (ING).* Ph.D. thesis, Stanford University Department of Linguistics, Palo Alto, CA, 2005.

Christy, A. The effects of explicit knowledge of and implicit attitudes about race on adult perceptions of children's speech., 2010. M.A. Thesis, Department of Speech-Language-Hearing Sciences, University of Minnesota.

Clopper, C. G. *Linguistic Experience and the Perceptual Classification of Dialect Variation.* Ph.D. thesis, Indiana University, 2004.

Coetzee, A. and McGowan, K. B. Allophonic cues to syllabification. In *CUNY Phonology Forum Conference on the Syllable*. New York, 2008.

Creel, S. C., Aslin, R. N. and Tanenhaus, M. K. Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106:633–664, 2008.

Croot, K. The emergent paradigm in laboratory phonology: Phonological categories and statistical generalisation in cutler, beckman and edwards, frisch and bréa-spahn, kapatsinski, and walter. *Journal of Laboratory Phonology*, 1(2):415–424, 2010.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K. and Hogan, E. M. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16(5/6):507–534, 2001.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C. and Mehler, J. Epenthetic vowels in japanese: a perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, 25:1568—1578, 1999.

Eberhard, K., Spivey-Knowlton, M.J., S. J. and Tanenhaus, M. Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24:409–436, 1995.

Foulkes, P. Exploring social-indexical variation: a long past but a short history. laboratory phonology. *Journal of Laboratory Phonology*, 1, 2010.

Foulkes, P. and Docherty, G. The social life of phonetics and phonology. *Journal of Phonetics*, 34:409–438, 2006. The social life of phonetics and phonology.

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

Ganong, W. F. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6:110–125, 1980.

Gerstman, L. Classification of self-normalized vowels. *Audio and Electroacoustics, IEEE Transactions on*, 16(1):78 – 80, 1968. ISSN 0018-9278.

Gick, B. and Derrick, D. Aero-tactile integration in speech perception. *Nature*, 462:502–504, 2009.

Goldinger, S. A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of 2007 International Congress on Phonetic Sciences.*, pages 49–54. Saarbrucken, Germany., 2007.

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

Green, K., Kuhl, P., Meltzoff, A. and Stevens, E. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the mcgurk effect. *Attention, Perception, & Psychophysics*, 50:524–536, 1991. ISSN 1943-3921. 10.3758/BF03207536.

Greenwald, A. G., McGhee, D. E. and Schwartz, J. L. K. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74:1464–1480, 1998.

Harris, Z. S. *Structural Linguistics.* University of Chicago Press, Illinois, U.S.A., 1951.

Hawkins, S. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3-4):373–405, 2003.

Hay, J. and Drager, K. Stuffed toys and speech perception. *Linguistics*, 48(4):865–892, 2010.

Hay, J., Nolan, A. and Drager, K. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, 23(3):351–379, 2006a.

Hay, J., Warren, P. and Drager, K. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34:458–484, 2006b.

Hintzman, D. L. Schema-abstraction in a multiple- trace memory model. *Psychological Review*, 93:411– 427, 1986.

Huettig, F., Rommers, J. and Meyer, A. S. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2):151 – 171, 2011. ISSN 0001-6918. doi:DOI: 10.1016/j.actpsy.2010.11.003. Visual search and visual world: Interactions among visual attention, language, and working memory.

Irvine, J. When talk isn't cheap: language and political economy. *American Ethnologist*, 16(2):248–267, 1989.

Irvine, J. T. and Gal, S. Language ideology and linguistic differentiation. In Kroskrity, P., editor, *Regimes of Language*, pages 35–83. School of American Research Press, Santa Fe, NM, 2000.

Ito, T. A., Thompson, E. and Cacioppo, J. T. Tracking the timecourse of social perception: The effects of racial cues on event-related brain potentials. *Personality and Social Psychology Bulletin*, 30:1267–1280, 2004.

Johnson, K. Speech perception without speaker normalization: An exemplar model. In Johnson, K. and Mullennix, J. W., editors, *Talker Variability in Speech Processing.*, pages 145–165. Academic Press, San Diego, 1997.

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

Joos, M. Acoustic phonetics. *Language*, 24(2):5–136, 1948. ISSN 00978507.

Jusczyk, P. W. and Luce, P. A. Speech perception and spoken word recognition: Past and present. *Ear and Hearing*, 23(1):2–40, 2002.

Kennedy, K. M., Hope, K. and Raz, N. Lifespan adult faces: Norms for age, familiarity, memorability, mood, and picture quality. *Experimental Aging Research*, 35(2):268–275, 2009.

Kirkpatrick, A. and Zhichang, X. Chinese pragmatic norms and 'china english'. *World Englishes*, 21(2):269–279, 2002. ISSN 1467-971X. doi:10.1111/1467-971X.00247.

Klatt, D. H. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–312, 1979.

Kucera, H. and Francis, W. N. *Computational analysis of present-day American English.* Brown University Press., Providence, 1967.

Labov, W. *Principles of linguistic change: Internal factors.* Wiley-Blackwell, 1994.

Labov, W. A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34(4):500 – 515, 2006. ISSN 0095-4470. doi:DOI: 10.1016/j.wocn.2006.05.002. Modelling Sociophonetic Variation.

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

Lambert, W. E., H. R. G. R. C. . F. S. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, 3:44–51, 1960.

Liberman, A. M. *Speech: A Special Code.* MIT Press, Cambridge, MA, 1996.

Liberman, A. M., Delattre, P. C. and Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 65:497–516, 1952.

Licklider, J. C. R. and Miller, G. A. The perception of speech. In Stevens, S. S., editor, *Handbook of experimental psychology.*, pages 1040–1074. Wiley, New York, 1951.

Lindblom, B. Explaining phonetic variation: A sketch of the h&h theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers, The Netherlands, 1990.

Lippi-Green, R. *English with an Accent:Language,Ideology,and Discrimination in the United States.* Routledge, London, 1997.

Lisker, L., . A. A. S. Some effects of context on voice onset time in english stops. *Language and Speech*, 10(1):1–28, 1967.

Lisker, L. and Abramson, A. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20:384–422, 1964.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N. and Dahan, D. Spoken word recognition in the visual world paradigm reflects the structure of the entire lexicon. In Hahn, M. and Stoness, S., editors, *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, pages 331–336. Erlbaum Associates, Mahwah, NJ, 1999.

Mann, V. A. Influence of preceding liquid on stop consonant perception. *Perception & Psychophysics*, 28:407–412, 1980.

McGurk, H. and MacDonald, J. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

McLennan, C. T. and Luce, P. A. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31:306–321, 2005.

Miller, J. D. Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, 85(5):2114–2134, 1989. doi:10.1121/1.397862.

Miller, J. L. and Liberman, A. M. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25(6):457–465, 1979.

Milroy, L. and Gordon, M. *Sociolinguistics: Method and Interpretation.* Wiley-Blackwell, Oxford, U.K., 2003.

Minear, M. and Park, D. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments and Computers.*, 36:630–633, 2004.

Miyawaki, K., Strange, W., R., V., Liberman, A., Jenkins, J. and Fujimura, O. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception and Psychophysics*, 18(5):331–340, 1975.

Munson, B. Levels of phonological abstraction and knowledge of socially motivated speech-sound variation: a review, a proposal, and a commentary on the papers by clopper, pierrehumbert, and tamati; drager; foulkes; mack; and smith, hall, and munson. *Journal of Laboratory Phonology*, 1:157–177, 2010.

Neuhauser, S. and Simpson, A. P. Imitated or authentic? listeners' judgements of foreign accents. In *ICPhS XVI*. International Congress of Phonetic Sciences, Saarbrücken, 2007.

Newman, M. and Wu, A. "do you sound asian when you speak english?" racial identification and voice in chinese and korean americans' english. *American Speech*, 86(2):152–178, 2011.

Niedzielski, N. Acoustic analysis and language attitudes in detroit. In Meyerhoff, M., editor, *(N)Waves and means: University of Pennsylvania working papers in lingusitics*, volume 3. University of Pennsylvania Press, Philadelphia, 1995.

Niedzielski, N. *The effect of social information on the phonetic perception of sociolinguistic variables.* Ph.D. thesis, University of California, Santa Barbara, 1997.

Niedzielski, N. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1):62–85, 1999.

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

Nosofsky, R. M. and Zaki, S. R. Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology*, 28(5):924–940, 2002.

Olive, J., Greenwood, A. and Coleman, J. *Acoustics of American English Speech.* Springer-Verlag, New York, 1993.

Peterson, G. E. and Barney, H. L. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952. doi: 10.1121/1.1906875.

Pierrehumbert, J. B. Exemplar dynamics: Word frequency, lenition and contrast. In *In*, pages 137–157. John Benjamins, 2001.

Pierrehumbert, J. B. The next toolkit. *Journal of Phonetics*, 34:516–530, 2006.

Port, R. F. Rich phonology: Some background material. http://www.cs.indiana.edu/ port/HDphonol/HDphonology.supporting.materials.html, 2008.

Potter, R. K. and Steinberg, J. C. Toward the specification of speech. *The Journal of the Acoustical Society of America*, 22(6):807–820, 1950. doi:10.1121/1.1906694.

Qiong, H. X. Why china english should stand alongside british, american, and the other. *English Today*, 20(02):26–33, 2004. doi:10.1017/S0266078404002056.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

Repp, B. B. The role of psychophysics in understanding speech perception. 1987.

Rubin, D. L. Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, 33(4):511–531, 1992.

Silverstein, M. Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4):193 – 229, 2003. ISSN 0271-5309. doi:DOI: 10.1016/S0271-5309(03)00013-2. Words and Beyond: Linguistic and Semiotic Studies of Sociocultural Order.

Squires, L. *Sociolinguistic Priming and the Perception of Agreement Variation: Testing Predictions of Exemplar-Theoretic Grammar*. Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2011.

Staum Casasanto, L. Does social information influence sentence processing? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Washington, D.C., 2008.

Staum Casasanto, L. *Experimental Investigations of Sociolinguistic Knowledge*. Ph.D. thesis, Stanford University Department of Linguistics, 2009a.

Staum Casasanto, L. How do listeners represent sociolinguistic knowledge? In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Amsterdam, NL, 2009b.

Staum Casasanto, L. What do listeners know about sociolinguistic variation? *Penn Working Papers in Linguistics: Selected papers from NWAV 37.*, 15(2), 2010.

Strand, E. A. *Gender Stereotype Effects in Speech Processing.* Ph.D. thesis, The Ohio State University, 2000.

Tang, C. and van Heuven, V. J. Mutual intelligibility of chinese dialects experimentally tested. *Lingua*, 119(5):709 – 732, 2009. ISSN 0024-3841. doi:DOI: 10.1016/j.lingua.2008.10.001.

Van Engen, K. and Bradlow, A. R. Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America*, 121(1):519–526, 2007.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M. and Bradlow, A. R. The wildcat corpus of native- and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 2010.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S. and Munhall, K. G. Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, 60(6):926–940, 1998.

Whalen, D. H. Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, 35:49–64, 1984.

Willis, C. Perception of vowel phonemes in fort erie, ontario, canada, and buffalo, new york: An application of synthetic vowel categorization tests to dialectology. *J Speech Hear Res*, 15(2):246–255, 1972.