Socioindexical expectation and speech perception in noise: experienced and inexperienced listeners

Kevin B. McGowan
Rice University

November 15, 2012

Address for correspondence:
Kevin B. McGowan
Department of Linguistics, MS 23
Rice University
P.O. Box 1892
Houston, TX 77251-1892
U.S.A.

Acknowledgements:
This paper has benefited from insightful discussions with Pam Beddor, Steve Abney, Ben Munson, Robin Queen, Julie Boland, Andries Coetzee, Nancy Niedzielski, Anna Babel, Lauren Squires, Susan Lin, Anthony Brasher, the students in my Spring 2012 speech perception seminar, and audiences at the Linguistics Society of America, the Acoustical Society of America and the University of Texas at Austin. I would like to express my sincere thanks to all of the participants who took part in this study and to Keith Johnson for graciously hosting a portion of this research in the UC Berkeley phonology lab.

Abstract

A growing collection of results in sociophonetic speech perception has shown that listeners use cues to the social identity of an otherwise unknown speaker to impose structure upon phonetic detail. Socioindexical expectations activate expectations of social category, rather than individual identity, and have been shown to exert influence on a wide range of tasks. The existence of these effects has generally been interpreted to imply the existence of (1) listener experience with the phonetic cues, (2) knowledge of social categories, and (3) an accurate linkage of these two classes of information. This implied knowledge is, in turn, interpreted as evidence for exemplar theories of speech perception and their stored episodic memory traces. This paper investigates the extent to which accurate, neutral, and inaccurate socioindexal primes can influence listeners' ability to transcribe accented speech in noise. Results are obtained for populations of both experienced and inexperienced listeners. Contrary to a strict exemplar hypothesis, both groups transcribe more accurately when provided with an accurate socioindexical prime. However, patterning of the neutral and inaccurate primes suggests the need for a more nuanced interpretation of socioindexical perception results with a clear role for both detailed experience and stereotype.

Keywords: speech perception, sociophonetics, matched guise, perception in noise

INTRODUCTION

There is mounting evidence from laboratory phonology and sociophonetics demonstrating that socioindexical expectations can influence listener performance on a variety of behavioral measures (**?**). Manipulating listeners' beliefs about the age, gender, sexual orientation, race, etc. of the talker can lead to behavioral responses which suggest that listeners process the speech signal differently given different social expectations. Perception of particular phonetic cues can be altered in response to a primed social group –in other words, listeners appear to dynamically alter the attentional weights associated with particular phonetic cues in response to the manipulated social expectations (**?**). This alteration has been shown to occur both with cues that reflect actual usage (e.g. **??**) and stereotypical usage (e.g. **?**). Socioindexical perception effects suggest that speech perception proceeds, not by winnowing away noise to arrive at a core, intended signal, but by exploiting real patterns of informative variation to impose structure upon the phonetic signal.

This position, that systematic patterns of variation provide essential support for listeners, is certainly not without precedent. **?** finds that "variation is both necessary and beneficial" for English listeners overcoming a gross categorical mismatch in French-accented patterns of voice onset time (what **?** classifies as the single category assimilation of two variants). **?** demonstrates that listeners are sensitive to subcategorical acoustic mismatches below the level of experimenter awareness. The gating task of **?** finds that listeners hearing portions of CV and C$\tilde{\text{V}}$ stimuli report hearing a CVN, rather than CVC, target when vowel nasalization is present. **?** report eye-tracking evidence that coarticulatory cues speed the time course of lexical activation –with listeners in a forced-choice visual world task taking immediate advantage of available cues to nasalization. **?**, citing decades of research on the perceptual consequences of coarticulation, argues that coarticulation provides listeners with informative variation which gives structure to the acoustic signal and simplifies the task of perception. Individual listeners assign different weights to particular acoustic cues consistent with their experience of the variability and usefulness of those cues in understanding speech. Allophonic and coarticulatory variation, then, have been shown to both support and enhance perception and lexical activation. The position taken in this paper is that patterns of socioindexical variation should prove to be similarly useful for listeners.

If listeners use patterns of socially-informative variation available in the speech stimulus and if this invokes a set of specific phonetic expectations that covary with contextual cues then manipulating social expectations should be able to *enhance*, not merely alter, perception of an acoustic stimulus. **?** have shown that listener experience with a particular accent –Long Island English– predicts the usefulness of semantic priming by voices with that accent. Similarly, **?** used a cross-language and cross-dialect priming task to demonstrate that variable sociophonetic cues can facilitate translation priming. In this paper, I will present evidence from an experiment in which listeners were asked to transcribe Chinese-accented speech in noise. The results of this experiment suggest that manipulation of social expectations about a given talker can, in fact, enhance listeners' ability to disambiguate accented speech in a noisy signal. Listeners are more accurate when transcribing Chinese-accented speech in noise when shown an ostensible speaker with a Chinese face than when presented with a Caucasian face. This performance is compared to results in a control condition in which the guise is a schematic human silhouette.

The present result provides support for the usefulness of patterns of socioindexical variation and therefore provides support for a model of speech perception in which listeners' linguistic representations encode this variation. This finding is inconsistent with phonological models and theories of speech perception in which listeners are presumed to abstract away from detailed knowledge of the signal to an idealized or underspecified form.

Particularly, this outcome provides support for models in which listeners can link social category information to particular acoustic cues. While exemplar models certainly have both of these affordances (**???**), it does not necessarily follow that listeners store detailed episodic traces of experience and draw upon these experiences during perception. One can imagine, for example, extending the double-weak model (**??**) to include a set of social cues whose settings alter the weights assigned to particular acoustic cues to segmental and prosodic identity. Indeed, these affordances are consistent with any model of speech perception in which it is conceded that listeners are capable of learning. This could, in principle, include storage-intensive K-nearest-neighbor type algorithms, of which exemplar models are a particular instantiation, or simple Bayesian models which presume the storage of no training data at all (**??**). Exemplar models are presently an influential, if not *the* dominant, model for speech perception and lexical activation. Furthermore, the social affordances of exemplar models motivated –and are typically claimed to be supported by– the active research, over the last decade, in socioindexical speech perception. It is therefore essential that we devise experiments to test the specific predictions of these models to accumulate converging evidence or to develop alternative models with greater explanatory power.

Perhaps one means of testing the particular claims of exemplar models in a behavioral task is to control for listener experience in such a way that one group of listeners has demonstrated detailed knowledge of a particular speech variety and a second group has demonstrated a lack of this knowledge. In the experiment described here, this is implemented by dividing listeners into either an 'experienced' or 'inexperienced' group. The result of this manipulation is surprising in that experienced and inexperienced listeners appear to benefit approximately equally from the presence of accurate socioindexical information about the speaker. This issue is explored further in the Discussion and Conclusion sections.

Ideology or Experience?

In the model described so far, listeners use available cues to talker social category to simplify the task of perception by invoking either attentional cue weights or patterns of phonetic experience consistent with that social category. It is necessary, at this point, to briefly consider one alternative hypothesis that the results of socioindexical speech perception experiments may be best understood as resulting from listeners' language ideologies. The matched guise task presented here is, in spirit, a replication of **?**. One task in Rubin's study 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 cloze test –although this difference 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), students' firm belief that foreign graduate students will be incomprehensible will still result in perceptions of accentedness and reduced perceptibility.

This interpretation is later endorsed and extended by **?**. Lippi-Green introduces the concept of 'communicative burden' which is conceptually similar, though at a social level, to the H&H theory of speech perception (**?**). 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 this view, negative social bias leads listeners to choose to expend less energy decoding the acoustic signal, resolving ambiguous segments, etc.

Listeners in Rubin's 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." (**?**, 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.

Such an interpretation apparently stands in stark contrast to a model in which listeners are innocently using social knowledge (including social stereotype) to structure and interpret the acoustic signal. However, it may be possible to reconcile these positions by noting that Rubin's participants faired more poorly in the incongruous or mismatched condition. It may be the case, although it is difficult to discern from Rubin's results, that listeners in the congruous condition (Caucasian face paired with Standard American English voice) either responded at ceiling or enjoyed enhanced perception due to the presence of the supporting socioindexical prime. Listeners in the incongruous condition, on the other hand, were led to believe the graduate instructor was Asian. For them, the Chinese face may well have primed inaccurate attentional weights or stored episodic traces consistent with Chinese-accented English. One can straight-forwardly imagine this mismatch between social expectation and acoustic signal resulting in degraded performance.

To test this hypothesis, the current design employs, alongside congruous and incongruous conditions, a control condition in which listeners are presented with neither accurate nor inaccurate social information about the speaker. To ensure that listeners in this guise are still motivated to think about the identity of the speaker but with no cues to that identity, and to keep the visual and auditory structure of the trials as similar as possible across guise conditions, this control condition was implemented using a simple, geometric silhouette[1].

Rubin's is not the only evidence presented in the literature for the ideology hypothesis. In her classic study, **?** played recordings of a native Detroit, Michigan speaker for two groups of Detroit listeners. There is a widely held ideology among Michigan speakers that their variety of English is 'unaccented' or identical to Standard American English (**??**). Though both groups of Detroiters heard identical recordings of Detroit-accented speech, one group was led, in an incongruous condition, to believe the speaker was Canadian while the second group was led, in a congruous condition, 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 Canadian perceived more Canadian Raising than listeners who –correctly– believed the speaker to be from Detroit. Listeners who believed the speaker to be a fellow Detroiter were also less likely –though not significantly– to choose resynthesized vowels consistent with Detroit speakers' general participation in the Northern Cities Chain Shift (**?**). Instead, these listeners, in both the Canadian and Detroiter conditions, showed a preference for resynthesized vowels unlike the Northern Cities tokens they'd heard and

[1] A number of more realistic silhouettes were tested during pilot phases of this experiment, but participants showed a remarkable and somewhat unexpected ability to extract an accurate percept of 'Chinese' or 'Asian' from these visual stimuli.

closer to a model of Standard American English. While Canadians do not, in fact, participate in the Northern Cities shift, Detroiters most certainly do (**?**). So these Detroit listeners consistently made incorrect vowel identification decisions in both the congruent and incongruent conditions. Niedzielski interprets this result as evidence that language ideologies can interfere with listeners' access to fine phonetic detail during perception. This finding would appear to suggest, again, that the opposite of the proposed process is at work. Listeners use social expectation to enhance and give structure to the phonetic signal, however, it is not clear that simplifying and structuring the acoustic signal necessarily always means providing greater access to fine phonetic detail.

As also reported in **?**, **?** presented Swedish listeners who were highly proficient in American English with synthesized sVt continua. When instructions were provided in Swedish, listeners could not distinguish the *set*/*sat* portions of the continuum. When instructions were provided in English, however, this contrast was reliably percieved. Listeners' contextually motivated expectations suggested alternate listening strategies. Strategies which, in real world listening situations outside the speech perception laboratory, may greatly simplify the task of perception by reducing the set of phonetic features that must be attended to in real time to those which have a high probability of being informative. Niedzielski's listeners, from this perspective, are not attending to fine phonetic details beyond those required to ascertain category membership because they have not previously experienced a need to. These features of Detroit speech had not, as Niedzielski points out in her conclusion, risen to the level of consciousness among Detroiters.

Quantifying listener experience

? use essentially the same vowel selection task as **?**. Niedzielski's 'Michigan' and 'Canada' labels were replaced with 'New Zealand' and 'Australia'. Listeners in the Australian condition were more likely to perceive the stigmatized and socially-salient vowel /ɪ/ as higher and more front while listeners in the New Zealander condition perceived this vowel as more centralized. The same listeners, however, did not differ in their perceptions of the non-stigmatized, less-salient /ɛ/. As also discussed by Drager in **?**, this result suggests that perception effects are correlated with listener consciousness of the variant. The more stereotypical or available for metapragmatic discourse a variant is, the more likely it is that socioindexical priming of the appropriate social category will influence perception of that variable.

Hay, Nolan, and Drager's proposed link between social salience and socioindexical activation suggests an empirical method of dividing listeners into 'experienced' and 'inexperienced' groups of listeners for the purposes of a sociophonetic perception task. Listeners can be asked to demonstrate their ability to detect phonetic variables in a particular variety both above and below the level of social-salience for that variety. Performance on this task can be used to quantify each listener's degree of experience with that variety and establish a metric of that experience. For the purposes of the present study, participants, after completing the primary speech perception in noise task, were asked to participate in an accent identification task. This task gauged listeners' ability to distinguish authentic Chinese-accented English from a variety of other accents which included imitated Chinese.

The use of imitated Chinese accents was inspired by **?**, 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. It was 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 judgements. 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. In particular, listeners with authentic experience with a variety should be more capable of recognizing markers and perhaps indicators (**?**) of that variety than listeners with no direct or limited direct experience with that authentic variety. The 'experienced' and 'inexperienced' labels assigned using this task will be used in the analysis and discussion of the present transcription experiment; for a more thorough treatment of the accent identification task, see **?**.

Overview

As noted above, the present experiment is intended to be essentially a replication of **?** to address both lingering questions about the interpretation of that result and the advances made in sociophonetic perception since it was published. 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. In the experiment reported here, listeners hear recordings of Chinese-accented English and different faces are displayed to shift socioindexical expectations during transcription. However, the alignment of congruous and incongruous face and voice pairs has been inverted from Rubin's design. Specifically, in **?**, 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 **?**, expectation of a foreign accent can have a facilitatory effect on the understanding of accented speech. 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. Also as discussed above, 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. Listeners are grouped by experience level with authentic Chinese-accented English to test the hypothesis that socioindexical effects in perception provide strong evidence in favor of exemplar theories of speech perception.

METHODOLOGY

Stimuli

Stimulus materials consisted of 30 pairs of high and low predictability sentences originally developed by **?** 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 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 (**?**) 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 full set of sentences used are listed in appendix **??**.

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 (**?**) 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.

The actual target norms for L2 English speakers in China have traditionally been British rather than American English (**?**). Though there may be a shift currently underway 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.

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 (**??**) 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.

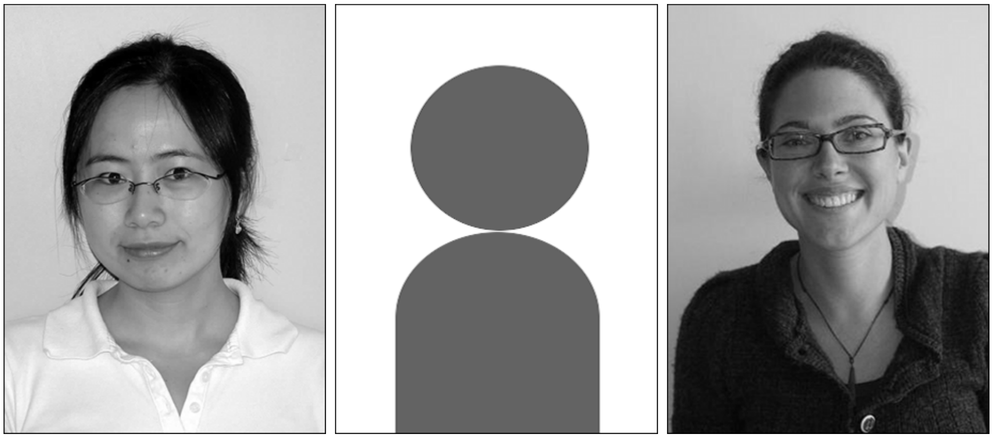

Figure 1: **Faces used in the transcription experiment**

One of three images was presented to listeners to establish these socioindexical expectations; these faces are shown in figure **??**. 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.

0.1 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.

0.1.1 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. Seven participants were excluded prior to data analysis. 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.

0.1.2 Experienced Listeners

Identifying a sufficiently large experienced population of Chinese-English listeners at the original University of Michigan research site 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 would not, in fact, represent an experienced population.

Thirty-one Heritage Mandarin-speaking undergraduate students from the University of California, Berkeley participated in exchange for $15.00. Two participants were removed prior to data analysis. In addition to the these two discarded subjects, 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.

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 used the same computers in the University of California, Berkeley's phonology laboratory. 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 guise 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?

Predictions

If exemplar theories of speech perception (e.g. **?**) 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 accent identification task 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.

RESULTS

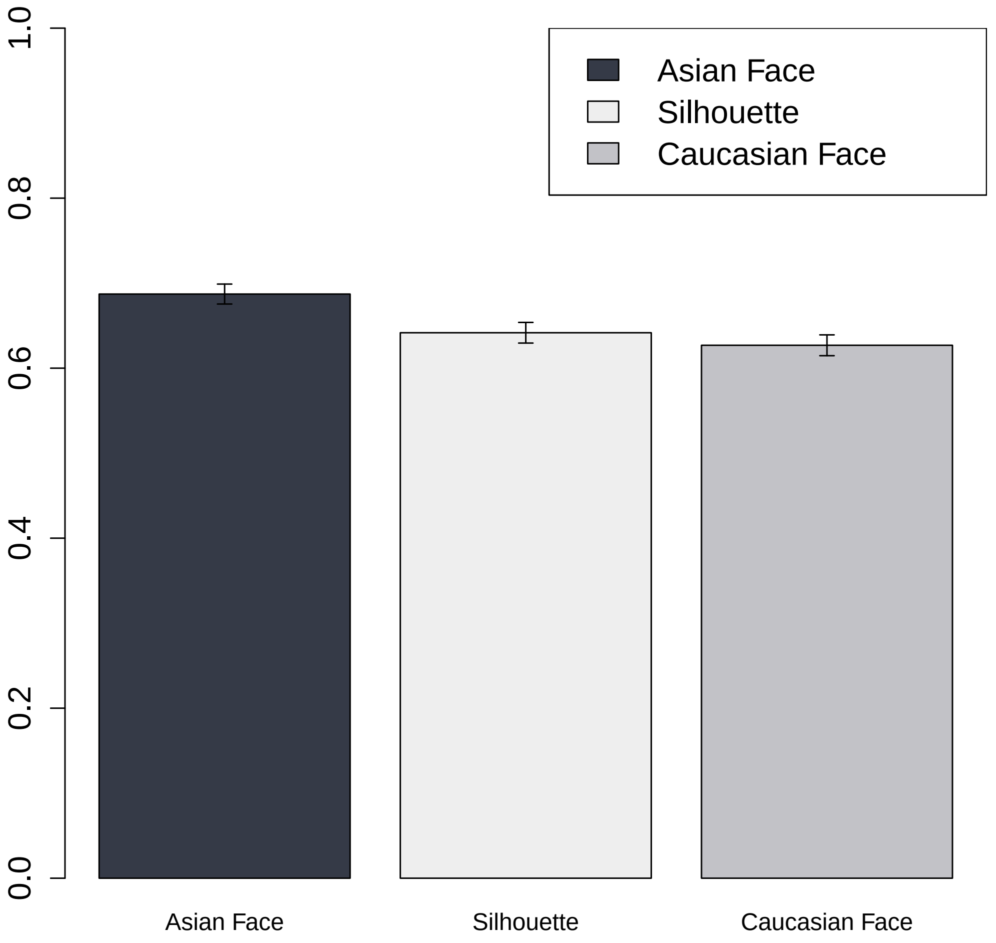

Figure 2: **All listeners: proportion correct target word responses for combined inexperienced and experienced listeners**

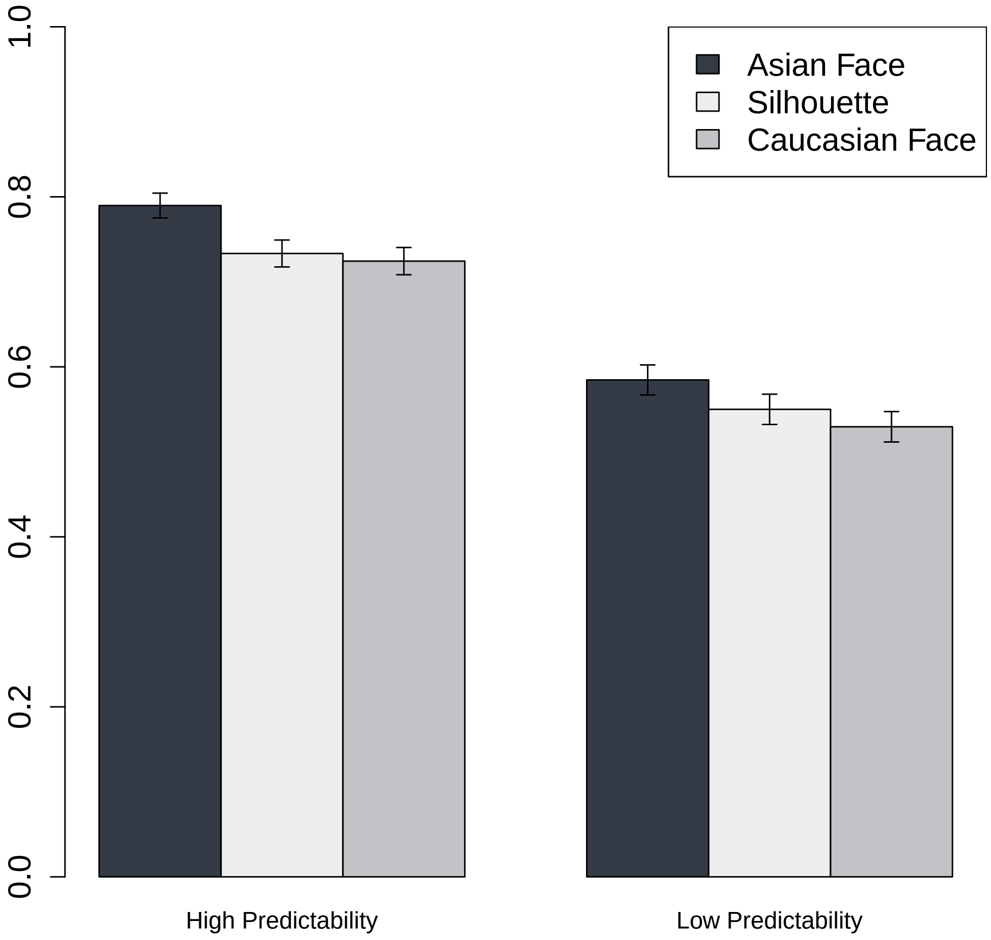

Figure 3: **All listeners: proportion correct target word responses for combined inexperienced and experienced listeners separated by Predictability**

	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 1: **Correct responses by Face, Predictability and Experience level**

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 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 experiment. 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 (**?**) and the packages `lme4` (**?**) and `languageR` (**?**).

Figure **??** 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 being significantly more accurate than transcription in the Caucasian condition. Table **??** 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 2: **The interaction of Face and Experience in terms of correct responses by all listeners**

Figure **??** 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 **??**). 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: **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

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

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

Inexperienced Listener Results

A further planned analysis step was 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 **??** 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 **??**, 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, 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.

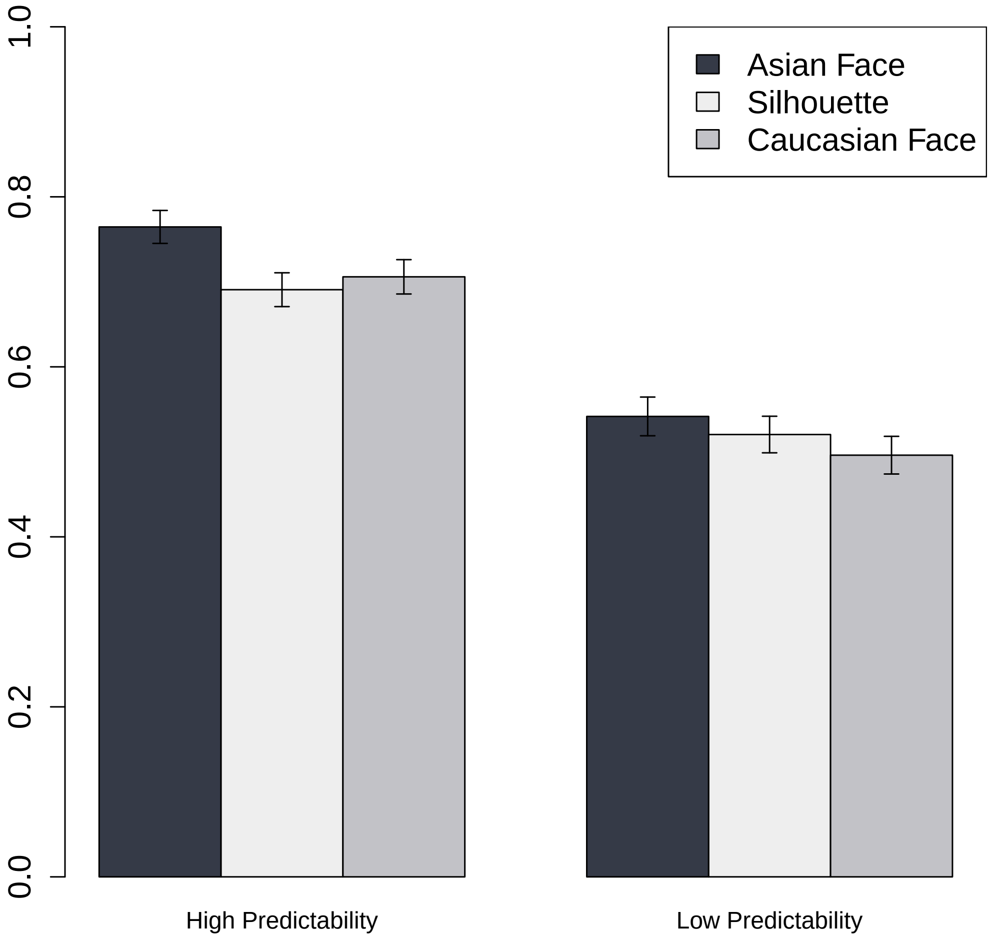

Figure 4: **Inexperienced Listeners: proportion correct target word responses**

Reported in table **??**, 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 ($\beta = -0.2727, p = 0.071$) and the difference between the Asian and Caucasian Face conditions 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.

Experienced Listener Results

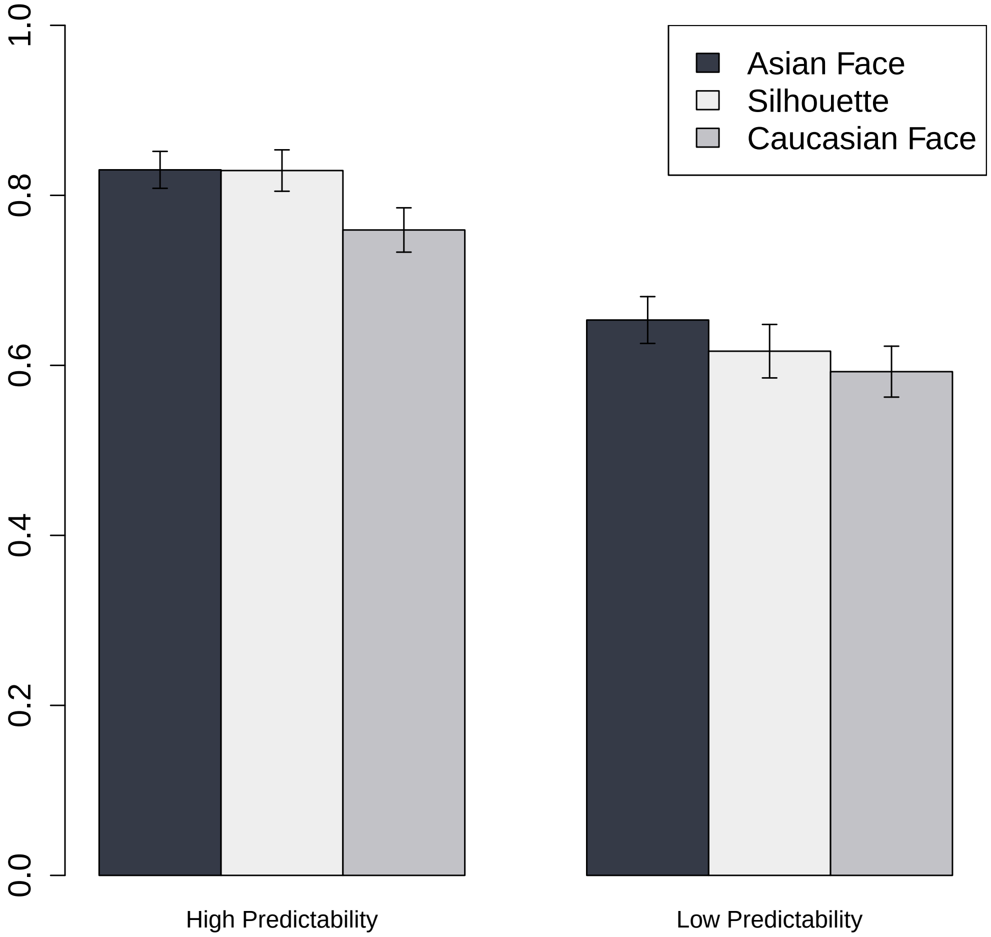

Figure 5: **Experienced Listeners: proportion correct target word responses**

Figure **??** 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 between the University of Michigan and UC Berkeley undergraduate populations. This possibility will be explored further in the Discussion.

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

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 function. Face, Predictability and the interaction of Face and Predictability were once again included as fixed effects in the model. Table **??** 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.

Indeed, when we average across the high and low predictability sentences and consider, as we did for the inexperienced listeners, a simplified model with only Face included as a fixed effect and Subject and Target word as random effects, there is a main effect of Face for experienced listeners ($\beta = 0.0428, p = 0.0428$).

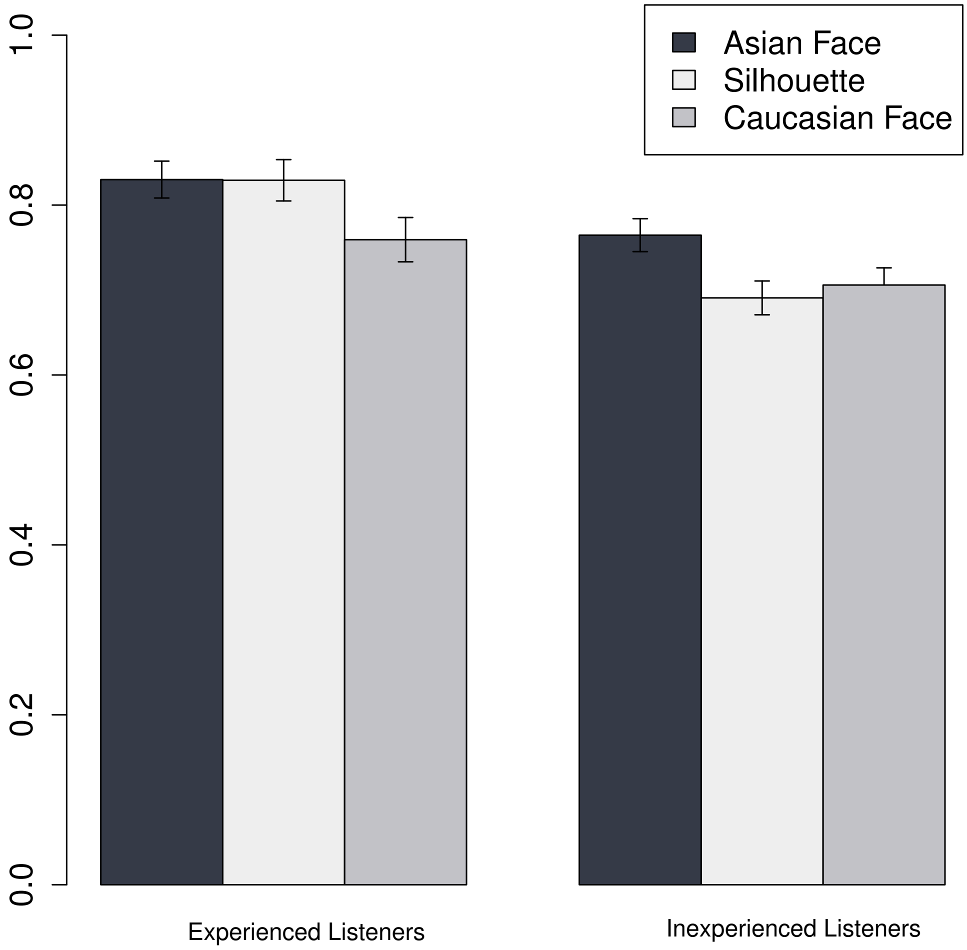

Figure 6: **Combined Experienced & Inexperienced Listeners: high predictability**

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 congruent with the voice they are listening to. However, it is not at all clear that facilitiation is the best description of the differences between guise conditions observed in these results.

The fact that the control 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 for inexperienced listeners.

Figure **??** combines the high predictability conditions for experienced and inexperienced listeners to facilitate comparison of the control silhouette condition. With the reference level of the Face variable changed to the 'Silhouette' level, silhouette is not significantly different from Caucasian face for experienced listeners ($\beta = -0.53, p = 0.0696$). However, this condition does clearly appear to cluster with the Asian face condition –in contrast to the Silhouette results in the Inexperienced condition, which cluster with the Caucasian face condition and are 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; this could easily have influenced their default assumptions about the identity of the speaker in the silhouette condition. Bearing that caveat in mind, it does appear that the manipulation of social information results in enhancement in the Asian guise for inexperienced listeners condition and inhibition in the Caucasian guise for experienced listeners.

Contrary to prediction, the magnitude difference in accuracy listeners demonstrate when seeing the Asian face and hearing Chinese-accented English is greater for High Predictability sentences than for Low Predictability. The original hypothesis had been that socioindexical expectation would facilitate perception of the most difficult signals more than when the signal to noise ratio was higher. Perhaps socioindexical information is most useful when it can reinforce, and possibly clarify, an almost accessible signal and is less useful when that signal is too degraded. This aspect of the results seems surprising given the reliability of sociophonetic effects shown in the literature.

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 produces a fairly rhotacized vowel 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.

CONCLUSION

The results reported here are problematic for any theory of speech perception that presumes abstraction away from variation in the signal and toward an idealized representation. In this respect, the general predictions of exemplar models are clearly upheld. Other evidence in support of exemplar models of perception comes from the fact that 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.

If social expectation were exerting a strong bottom-up influence on socioindexically appropriate exemplars, 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 listener expectations about the identity of the speaker are manipulated as described here. This more online task should reveal the extent to which socioindexical manipulation can 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 **?** 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 a negative one.

Whether a socioindexical prime enhances or interferes with listeners' use of phonetic detail is a function of acoustic cue, social category, listener identify, and context. Exemplar accounts can model these relationships, but results so far do not motivate the storage of episodic traces to the exclusion of other possible explanations. A naive Bayes classifier, with no storage requirements beyond the current state of expectation probabilities, can also model this relationship. Nearey's double-weak model, extended to include relationships to social variables, could similarly model the present results. Finally, the findings here reveal that laboratory characterization of listeners' experience is both possible and necessary to adequately evaluate and advance an exemplar theory of speech perception.

A HIGH AND LOW PREDICTABILITY SENTENCES

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 6: **High/Low predictability sentence pairs from ?**