VOICE ONSET TIME IN THE BILINGUAL MIND: THE ROLE OF SPEAKER'S NATIVE LANGUAGE STATUS ON SEQUENTIAL BILINGUAL LISTENERS' PHONETIC PERCEPTION

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

By

Ke Lin, M.S.

Washington, DC
August 22, 2025

VOICE ONSET TIME IN THE BILINGUAL MIND: THE ROLE OF SPEAKER'S NATIVE LANGUAGE STATUS ON SEQUENTIAL BILINGUAL LISTENERS' PHONETIC PERCEPTION

Ke Lin, M.S.

Thesis Advisor: Elizabeth Zsiga, Ph.D.

Abstract

A growing body of work in sociolinguistics and speech processing research has demonstrated that speech perception is a dynamic process through which listeners actively leverage linguistic and social knowledge to interpret auditory input. While extensive evidence shows that social cues bias phonetic perception for native (L1) listeners, these effects remain underexplored for second-language (L2) listeners. Bilingual listeners, who constantly navigate both the social and phonological dimensions of two languages, provide unique opportunities to refine existing models of bilingual processing and sociophonetic perception.

This dissertation investigates how sequential bilingual listeners of Mandarin–English ($N = 42$; $mean\ age = 28.3$, $SD = 10.7$) and Russian–English ($N = 39$; $mean\ age = 34.3$, $SD = 11.8$) perceive voicing contrasts along voice onset time continua under varying talker guises (native versus non-native). Two complementary paradigms were employed: (1) an explicit perceptual task using a Visual Analogue Scale (VAS) to assess listeners' categorization of synthesized syllables, and (2) an implicit anticipatory eye movement (AEM) task powered by OpenFace 2.0 (Baltrušaitis et al., 2016) capturing real-time processing of lexical items. Additionally, measures of English oral proficiency and cognitive styles were collected to explore how individual differences modulate perceptual adjustments.

Results from the VAS indicate that bilingual listeners' baseline voicing category boundaries often diverged from monolingual norms, shifting generally toward English (L2) phonetic categories. Importantly, social guises reliably induced perceptual shifts, especially when the acoustic input is ambiguous. In the AEM task, Mandarin–English listeners exhibited reliable talker-identity effects in real time, whereas Russian–English listeners did not. Moreover, higher English proficiency correlated with more pronounced socially-induced category shifts, while listeners with stronger autism-aligning traits in the Communication subscale exhibited reduced sensitivity to social information, maintaining more L1-like phonetic boundaries.

These findings support an interactive, expectation-driven model of speech perception, highlighting bilingual listeners' flexible re-weighting of acoustic cues according to perceived talker identity. This socially driven perceptual tuning is further modulated by L2 proficiency and sociocognitive style, enriching existing bilingual processing models—the Unitary Language System hypothesis (Volterra and Taeschner, 1978) and the Double Phonemic Boundary hypothesis (Caramazza et al., 1973)—by integrating social-contextual and individual cognitive factors.

Index words: speech perception, sociophonetics, bilingualism, second-language (L2) processing, voice onset time, eye-tracking

Dedication

For the bilinguals who live between languages, hearing themselves anew in every accent.

For those who have struggled to be understood—and kept speaking anyway.

For the researchers who listen closely to the voices between worlds.

For the Chinese-speaking and Russian-speaking communities whose voices shaped this research, and for the part of myself still learning to listen.

Acknowledgments

Many efforts, circumstances, and generous people have made this dissertation possible—but I would like to begin with a concept that has quietly yet profoundly shaped both my research and my outlook on life: *ostranenie*. I first encountered the idea of *ostranenie* (остранение, ‘defamiliarization’) while reading Viktor Shklovsky’s essay “*Искусство как приём*” (*Art as Technique*) for Prof. Angela Brintlinger’s seminar at The Ohio State University. Shklovsky wrote:

> Целью искусства является дать ощущение вещи, как видение, а не как узнавание; приемом искусства является прием „остранения“ вещей и прием затрудненной формы, увеличивающий трудность и долготу восприятия, так как воспринимательный процесс в искусстве самоцелен и должен быть продлен......
>
> *The technique of art is to make objects ’unfamiliar,’ to make forms difficult, to increase the difficulty and length of perception because the process of perception is an aesthetic end in itself and must be prolonged.* (Shklovsky, 1917, 1965).[1]

This revelation stayed with me.

As a sociolinguist and transnational navigating American, Chinese, and Russian worlds, this notion became a framework for questioning the assumptions we carry

[1] Original text from Shklovsky (1917), retrieved from `https://www.opojaz.ru/manifests/kakpriem.html`. English translation from Shklovsky (1965), in Lee T. Lemon and Marion J. Reis (Eds.), *Russian Formalist Criticism: Four Essays* (Lincoln: University of Nebraska Press, 1965).

about languages and identity—a reminder that research is not only about what we discover, but how we look. In the years since, this idea has guided my choices: to explore new disciplines — moving from regional studies in Slavic and Eastern European languages and cultures to quantitative, experimental sociophonetics; to take up marathon running as an exercise in endurance and renewal; and to stay curious through every twist of this long dissertation journey. It taught me that growth often begins when we step back from the familiar and learn to see the world anew.

My journey began at The Ohio State University. As a newly minted double major in Russian and Linguistics, I entered a Ph.D. program in Slavic linguistics without knowing exactly where it would lead. My coursework and teaching responsibilities opened more questions than it answered, but a fellowship in my third year allowed me to explore courses beyond my department—an opportunity that quietly redirected my path. In Professor Don Winford's Introduction to Sociolinguistics, I encountered the three waves of sociolinguistics through the pioneering works of William Labov, Lesley Milroy, and Penny Eckert, whose research resonated deeply with my own lifelong questions about language, identity, and social meaning. Prof. Angela Brintlinger's seminar, where we read Shklovsky, encouraged me to connect literature and cultural theory with linguistics. Professor Morgan Liu's course on Culture and Politics in Central Asia inspired my first mixed-methods project surveying language attitudes toward non-native speakers in Russia.

These experiences taught me that detours enrich rather than delay and gave me the courage to reframe my academic goals at Georgetown University, where I found a home in sociophonetics, bilingualism, and psycholinguistics. Here I learned how to combine acoustic analysis with social meaning and cognitive processes. The community I encountered at Georgetown embodied the art of mentorship—guidance that is rigorous yet deeply human.

It is impossible to speak of this journey without honoring the mentors whose teaching, patience, and example carried this work from its earliest idea to its completion. I will begin with Prof. Jennifer Nycz, who introduced me to the tools essential to sociophonetic research and to the broader and more nuanced world of variationist sociolinguistics. Across four courses, she taught me to work with ELAN, Praat, and RStudio, and encouraged me to keep exploring independently—to attend workshops and conferences, experiment with emerging text-to-speech technologies and statistical modeling approaches newly adopted in our field, and evaluate them alongside traditional methods of analysis and transcription. She also inspired us to develop our own technical tutorials as our skills grew. Prof. Nycz has an exceptional gift for remembering her students' interests and connecting us with researchers who share them; through her, I had the chance to speak with experts in social speech perception and matched-guise design such as Kevin McGowan and Ksenia Gnevsheva during the earliest stages of brainstorming this dissertation.

Prof. Elizabeth Zsiga, my advisor, provided steady and unwavering support throughout every stage of this project. Though her expertise lies in theoretical phonetics and phonology, she wholeheartedly encouraged my turn toward experimental sociophonetics and championed my use of newer tools—such as PCIbex for online experiment design and OpenFace for gaze tracking— when they are not yet common in our department. As her research assistant on a project investigating prosodic variation in Štokavian dialects, she patiently taught me the theoretical foundations of prosody and guided me through both manual and automated pitch manipulation. Her mentorship combined intellectual rigor with remarkable generosity: she offered hands-on training, helped me navigate IRB and funding procedures, and consistently supported the value of interdisciplinary thinking. Her confidence in my work, especially during moments of uncertainty, gave me the direction and assurance I needed to continue growing as a researcher.

Prof. Lacey Wade became my mentor and co-architect. I first encountered her research during an LSA webinar on expectation-driven convergence, where she demonstrated how listeners unconsciously adopt dialectal features linked to a talker's perceived background—even when those features aren't actually present! When we later met at Georgetown's departmental talk, our shared interest in how social expectations shape language behaviors quickly grew into collaboration. Lacey tirelessly taught me how to use PCIbex for online experiments, listened to countless rounds of VOT continua, guided me in refining stimulus randomization and block design, and reviewed my posters and talks with remarkable generosity. Her curiosity, precision, and endless patience made our weekly check-ins something I truly looked forward to.

Prof. Sarah Phillips became my guide to eye-tracking and psycholinguistics. From our first meetings, she helped me clarify my early ideas and explained how gaze patterns can reveal subtle judgments and hesitations in speech processing. She taught me how to align eye-tracking data with auditory stimuli and offered detailed feedback on both my analyses and writing. Beyond the dissertation, Sarah consistently looked out for my growth—sharing job opportunities, reading funding applications, and encouraging me to think long-term about my research and career. Her mentorship has shaped not only the technical rigor of my work but also the kind of mentor I hope to become.

Prof. Lourdes Ortega exemplified mentorship grounded in understanding the diverse pathways of language learning and bilingualism research. I first heard her speak about second-language acquisition in contexts as varied as tourism, study abroad, and immigration, and later audited her course Second Language Acquisition and Bilingualism. She showed me that research is not only about mechanisms of learning but also about helping individuals make sense of their own language experiences—their goals, motivations, and trajectories. Her feedback pushed me to situate my findings within broader discussions of how linguistic and cultural backgrounds shape learning outcomes and self-perception, from differences in L1–L2 interaction to the social realities of being recognized—or not—as a native speaker. Her perspective reminded me that understanding language learning requires attending not just to cognition, but to lived experience.

I am also grateful to Professors Meg Montee, Alexandra Johnston, Cynthia Gordon, Marissa Fond, Kristin Rock, Heidi Getz, and Amir Zeldes, whose courses and mentorship have profoundly shaped my development as a linguist. Their expertise—spanning discourse analysis, statistics, psycholinguistics, natural language processing, and language assessment—enriched every stage of my doctoral journey and broadened the way I think about language as both data and lived experience.

Equally important were the members of my research team—from volunteer research assistants to technical collaborator—whose energy and dedication sustained the daily work of this project, both in the lab and in the field. Patrick Mahoney and Catrina Kellagan reached out in my final year eager to learn about linguistics and research, and soon found themselves scoring nearly a hundred elicited-imitation tasks, meeting weekly without credit or compensation—aside from a healthy dose of linguistic and statistical character-building. Even as freshmen, they witnessed the unglamorous and sometimes chaotic side of research: the experimental sessions that didn't record, the audio files that corrupted, the data that refused to align, and the collective relief when everything finally did. Linxuan (Roxy) Xu, originally a pilot participant, joined as a volunteer research assistant and transcribed interviews with remarkable care and consistency. Marek Vondrak with expertise in computer vision developed the processing pipeline that aligned OpenFace outputs with experimental timing and provided critical insights that shaped the analysis stage. His contributions were indispensable to the success of our new eye-tracking paradigm. Alexander Thoman helped recruit Russian volunteers and organize off-campus sessions, including one memorable recording day in a not-yet-opened beauty salon above Rashad Gourmet. I am also deeply grateful to the Mandarin-English and Russian-English bilingual participants who traveled, sat through tedious tasks, and shared their experiences; their patience, trust, and generosity form the heart of this dissertation.

Beyond my mentors and dissertation research team, friends and communities sustained me. Georgetown peers—Dan DeGenaro, Xiulin Yang, Hannah Fedder Williams, Ping Hei Yeung, Luis Daniel Acevedo Vélez, Claire Henderson, Abby Killam Jarvis, Yunie Ku, Kris Cook, Yi Luo, Jordan Mackenzie, Helen Dominic, Amber Hall, and Khaled Alharbi—offered camaraderie, steadfast support, and a sense of solidarity through both long nights and hard seasons. My extra-Georgetown linguist friends—Xin Gao, Jack Yuanfan Ying, Tamay Levy, Mackenzie Hope, Chen Zhou, Sreeparna Sarkar, Kei Kuwahara, and Stepan Mikhailov—became my co-working companions and emotional anchors across time zones. Church friends (Pastor David Tian, Sister Shanya, and Qi Zhang) reminded me of grace and purpose. Ballroom dance partners (special shoutouts to Kirk Sigmon and Paul Bohman) and rock-climbing partner (Brian Lee Park) taught me balance and the art of falling gracefully, both on and off the wall. And Sai, who connected with me through jam sessions and obscure Chinese films, brought laughter and ease into a time when I was often too consumed by work to enjoy life.

Training for my first full marathon—three months of waking at 5 a.m., running 10–15 km for short runs along the Mt. Vernon Trail at sunrise, and returning home to shower before starting work—was both a discipline and a joy. When I crossed the finish line on December 1, 2024, I thought I could do anything from then on, only to realize that what remained of my dissertation project was like running several marathons back-to-back. Yet the endurance and humility cultivated through running became essential. My family's love—expressed in late-night messages, my mother's Weee! orders, and my father driving across states to attend my defense—anchored me through every period of turbulence. Strength and care from God sustained me and renewed my sense of mission whenever I felt lost, especially during the 66-day "data-collection marathon," when participants came back-to-back and I often worked ten-hour days without stopping to eat—troubleshooting experiments, meeting sessions, processing data, and writing late into the night. Through it all, I was reminded of the verse, "Let us run with endurance the race that is set before us" (Hebrews 12:1), which became both prayer and promise in completing this work.

Through these relationships—with friends, family, and the quiet strength I've found in God—I have come to see that the triad of connection, growth, and purpose anchors both my personal and professional life: connecting with people across languages and cultures, growing through detours and setbacks, and finding purpose in research that humanizes rather than alienates.

Returning to Shklovsky, I now see his insight not only as an artistic philosophy but as a way of living and doing science. He reminds us that art's value lies in the act of perception itself—that by making the familiar strange, we experience the world more fully. Research, too, is valuable not only for what we discover, but for how we approach it: with curiosity, care, and ethical attention. This dissertation is my attempt to "make the stone stony" (Shklovsky, 1917/1965, p. 12) - to re-see language, identity, and perception and to share that renewed vision. To everyone who has walked, run, climbed, laughed, or prayed with me along this path—mentors, collaborators, participants, friends, and family—thank you. I look forward to carrying this spirit of defamiliarization, endurance, and connection into the next chapter of my journey.

Contents

Appendix

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Socially Modulated Speech Perception in L1 vs. L2 Contexts

Speech perception is not a passive decoding of acoustics, but a dynamic process shaped by both bottom-up input and top-down expectations. Listeners actively leverage their linguistic and social knowledge to interpret what they hear. A growing body of research demonstrates that beliefs about a talker's background can systematically influence how incoming speech sounds are categorized (Strand, 1999; Rubin, 1992; Hay and Drager, 2010; McGowan, 2015; Gnevsheva, 2018; Wade, 2022). For example, listeners' perceptions have been shown to shift based on a talker's perceived age (Koops et al., 2008), gender (Strand, 1999), regional dialect (Niedzielski, 1999; Hay et al., 2006; Hay and Drager, 2010; Wade, 2022), or ethnic background (Rubin, 1992; McGowan, 2015, 2016; Gnevsheva, 2018), even when the acoustic signal itself is identical. In a classic demonstration, (Lambert et al., 1960) presented the same voice in different language guises and found that listeners' evaluations of the speaker changed purely as a function of their beliefs about the speaker's identity. Such studies cement the view that what we hear is filtered through what (and whom) we expect.

Current literature in the field offers robust evidence that social information biases perception for monolingual or native (L1) listeners, yet far less is known about how these effects play out for second-language (L2) listeners. It is reasonable to assume, as (McGowan, 2016) suggests, that all listeners use "linguistic and social knowledge to impose structure upon the sensory events of speech" (p. 26). Nevertheless, L2 listeners could experience bias effects in ways that diverge from L1 listeners due to their different linguistic backgrounds and social orientations. In particular, L2 listeners bring the influence of their native phonological system to perceiving the L2, and they may hold distinct attitudes or category boundaries for what counts as "native-like" speech. These perceptual and attitudinal frameworks also shape how L2 listeners draw social boundaries, especially in distinguishing "in-group" from "out-group" talkers. For instance, from a native English listener's perspective, a fellow native speaker is an in-group talker, whereas a person speaking English with a foreign accent may be perceived as an out-group member. By contrast, an L2 English listener may classify talkers along multiple dimensions of proximity to a speaker, including linguistic similarity, social alignment, and language attitudes. I categorize the speaker in three ways: (1) an L1 English speaker, (2) an L2 English speaker sharing the listener's L1, and (3) an L2 English speaker who does not share the listener's L1.

These differing orientations mean that L2 listeners may not mirror L1 listeners in their social perceptual biases. For instance, an English learner from China might be especially attuned to a Chinese-accented English speaker (a peer L2 speaker) and could conceivably understand a Chinese-accented utterance more readily than an American listener would. On the other hand, the same L2 listener might also hold native English speech as the model of correctness, leading them to accommodate or perceive speech differently when they believe the talker is a native speaker. In short, L2 speech perception may be biased in ways inherently different from L1 perception, due to L2 listeners' unique blend of L1-based phonological processing and their distinct social framing of who is "us" vs. "them" in a given speech context.

The present work is motivated by this gap in understanding. We know that native listeners routinely integrate social information (e.g. a talker's age, race, L1) into speech processing, but we know relatively little about how the same social cues affect L2 listeners' perception. If L2 listeners do use social cues, do they use them to the same extent, and in the same ways, as native listeners? Or do differences in linguistic experience and social group alignment lead to different outcomes? Addressing these questions is important for building a comprehensive theory of speech perception. It can reveal whether the mechanisms demonstrated in L1 perception studies are universal to all human speech processing, or whether they interact with language experience in important ways. This dissertation particularly examines adult sequential bilinguals of Mandarin–English and Russian–English, two groups that acquired English as their L2. These groups were chosen because their L1s occupy opposite ends of the VOT continuum (as will be discussed in Chapter 2) in drawing boundaries for voicing contrast, allowing us to observe potential perceptual shifts in both directions along that continuum. By investigating how social information about a talker's identity (for example, cues suggesting the talker is a native vs. non-native English speaker) might alter these bilinguals' categorization of English speech sounds, this dissertation explores whether the biasing effects observed in monolingual listeners extend to bilingual listeners, and if so, how the listeners' L1 background might modulate those effects.

In the sections that follow, I review key aspects of social speech perception and bilingual phonetic processing that form the foundation of this study. Section 1.2 examines the role of race and social expectations in shaping perception. Section 1.3 discusses findings on L2 phonetic convergence and the link between perception and production in bilinguals. Section 1.4 introduces the theories of bilingual phonetic processing and the double boundary hypothesis, which motivates the current investigation of context-sensitive category shift. Finally, Section 1.5 presents the research questions and sets of hypotheses driving the present work.

1.2 The Role of Race and Social Expectations in Speech Perception

Among the various social factors, race and ethnicity have emerged as powerful modulators of speech perception in many L1 studies (Kutlu et al., 2022c,b; Kutlu, 2023). In multicultural societies, listeners often carry ingrained expectations linking race with language ability or accent, and these expectations can bias what they hear. In U.S. English settings, for example, Whiteness is commonly associated with being a native English speaker, whereas individuals of Asian descent are frequently presumed to be non-native speakers until proven otherwise. These racialized expectations form an implicit backdrop that listeners may apply when processing speech. Recent work by (Kutlu, 2023) discusses how such raciolinguistic perceptions lead listeners to mishear or judge speech differently based on the racial appearance of a talker, even when the speech itself is the same. In other words, listeners often conflate race with linguistic competence, and this effect can skew their perception of accent or intelligibility.

A seminal study by (Rubin, 1992) brought this issue to light in the context of U.S. university classrooms. Rubin was interested in why students often complained about the "accents" of Asian teaching assistants. Rather than examining the TAs' actual speech, Rubin cleverly manipulated students' impressions of who the TAs were. He played an identical recording of a lecture given by a native English speaker to two groups of students, but showed each group a different photograph of the supposed speaker: one photo depicted an Asian woman, the other a Caucasian (White) woman. The results were striking: students who believed the speaker was Asian rated the voice as more "heavily accented" and less comprehensible than students who thought they were listening to a White speaker. In reality, the audio was identical in both conditions. Rubin's finding demonstrated that racial bias alone – the mere sight of an Asian face – could create the illusion of a foreign accent and even impair comprehension. This study exposed how profoundly race-based expectations can shape one's auditory experience, effectively injecting an accent into the listener's mind even when acoustically there is none.

Subsequent research built on Rubins's (Rubin, 1992) pioneering work, examining not only the negative effects of a race–accent mismatch but also cases where matching a talker's race and accent can aid perception. (McGowan, 2015) showed that in difficult listening conditions, aligning the expected race of a talker with their accent improves listeners' comprehension. In McGowan's experiment, native English listeners heard sentences spoken in Chinese-accented English against noise. When these sentences were presented with a picture of an Asian face (congruent with the Chinese accent), listeners were significantly better at understanding the content than when the same sentences were paired with a White face. In essence, when the talker "looked" Asian, listeners' brains expected an Asian accent and coped better with it, whereas a mismatch between face and voice (an Asian accent coming from a White face) disrupted processing. This finding is often interpreted through the lens of reverse linguistic stereotyping and the so-called "mismatch effect". Listeners carry stereotypes about how certain groups speak, and when incoming cues violate those stereotypes (e.g., an accent that doesn't "match" the speaker's appearance), processing is disrupted.

Other studies have revealed more nuanced ways that racial expectations infiltrate perception. (Gnevsheva, 2018), for instance, examined how native English listeners in New Zealand rated the accentedness of audio clips when provided different visual or contextual information. Listeners heard English speech from L1 Korean, L1 German, and L1 New Zealand English speakers under three conditions: audio-only, audio+video, and video-only (no audio). Intriguingly, Gnevsheva found that when no audio was present (video-only), listeners still formed expectations about accent based on the talker's appearance, and that these expectations could lead to imagined accents. In the case of a Korean speaker, participants in the video-only condition expected a strong accent and later, when hearing the voice or reflecting on it, rated the Korean speaker as having a heavier accent than they actually did. Meanwhile, for a German speaker, seeing the talker reduced the perceived accent in the audio-visual condition (perhaps because the German speaker did not fit the listeners' stereotype of a heavily accented English speaker). These results demonstrate that listeners actively use racial and ethnic information to "fill in" linguistic qualities. They may imagine a thicker accent for someone who looks Asian – even in the absence of audio – and conversely might downplay an accent when the talker looks more familiar or "expected" to them. Racial cues, therefore, do not merely add a slight bias at the margins of perception; they can fundamentally alter what listeners think they hear, to the point of creating percepts that align with stereotypes rather than with the acoustic reality.

Notably, social expectations can be manipulated through explicit labels and symbols as well as through a talker's physical appearance. In a classic study on dialect perception, (Niedzielski, 1999) presented Detroit listeners with the same ambiguous vowel sounds but told them in one condition that the talker was from Michigan (their home region) and in another that the talker was from Canada. When listeners believed the talker was Canadian, they reported hearing the vowels as more "Canadian-sounding" (raised ou vowels, a feature associated with Canadian English), but when the same speech was labeled "Michigan," listeners did not report hearing the raised quality. Simply changing a written label ("CANADA" vs. "MICHIGAN") on the test materials was enough to shift how people categorized a vowel sound. Similarly, (Hay and Drager, 2010) famously showed that even subtle, non-human cues can prime listeners' expectations. They placed stuffed toy animals (a kangaroo vs. a kiwi bird) in plain view of participants during a speech perception task; when the kangaroo was visible, New Zealand listeners were biased to hear vowels as more Australian-like, whereas the kiwi toy cued more New Zealand-like vowel perceptions. Neither toy "speaks," of course, but each symbolically invoked a national stereotype that influenced perception. These findings drive home a critical point: speech perception is deeply social. Our brains do not process speech sounds in a vacuum – we constantly incorporate contextual clues about who is speaking, where they are from, and what social attributes they have, often without conscious awareness. Race is one of the most salient and consequential of these attributes in many settings, as it can trigger a cascade of associated beliefs (about accent, intelligence, status, etc.) that listeners then project onto the speech signal.

For the present research on bilingual L2 listeners' socially modulated speech perception, the role of race-based expectations is a central concern. If native English listeners imagine accents or adjust comprehension based on a talker's race, will L2 English listeners do the same? On one hand, L2 listeners might be less susceptible to some biases; for instance, a Chinese L2-English listener might not assume that an Asian-looking person with L2 background must have an accent, if the listener deems themself as proof that an L2 speaker can speak fluent English. On the other hand, L2 listeners might have their own set of biases – perhaps preferring the speech of "native-English-looking" individuals because they aspire to that pronunciation, or conversely, finding comfort and intelligibility in the speech of someone who shares their ethnicity. The prior L1 studies suggest a range of outcomes is possible. What they definitively illustrate is that socially driven perception is powerful: listeners can be "tricked" by their expectations, sometimes for the worse (as in (Rubin, 1992)'s study, where bias created comprehension problems) and sometimes in facilitating ways (as in (McGowan, 2015)'s congruent face case improving comprehension). The current study builds on this insight to examine whether similar effects extend to bilinguals by asking whether social cues about a speaker's racial and linguistic identity alter bilinguals' phonetic categorization in English. If L2 listeners show perceptual shifts patterned by talker identity—as monolinguals have—this would extend theories of socially modulated speech processing into the bilingual domain. If they do not, this may indicate that bilingual experience modulates or attenuates such biases.

1.3 L2 Phonetic Convergence and the Perception–Production Link

Listeners are not only perceivers but also potential imitators. A robust finding in phonetics is that people tend to converge toward the speech patterns of their interlocutors (often unconsciously) – a phenomenon known as phonetic accommodation (Giles et al., 1991; Giles and Powesland, 1997). This effect has been documented widely among native speakers, but it is particularly pertinent in second-language contexts. L2 speakers often exhibit a strong inclination to imitate or adopt the pronunciation features of native speakers when speaking the L2. In other words, they converge toward what they perceive as the target norm. For example, in a study of Polish learners of English, (Zając and Rojczyk, 2014) found that participants who shadowed English vowels tended to make their vowel durations more like a native British English model's when the model was a native speaker – but if the model speaker had a noticeable Polish accent, the learners did not converge in the same way (in fact, they sometimes diverged). This finding suggests that L2 speakers preferentially imitate what they believe to be the more native-like pronunciation. Even without explicit instructions to "sound like" the model, these learners shifted their production toward the native English speaker's vowels far more than toward a fellow L2 speaker's, effectively distancing themselves from the non-native model. Such behavior likely reflects both a cognitive inclination to align with a perceived standard and a social desire to fit in with native norms.

Social beliefs can amplify convergence effects. (Jiang and Kennison, 2022) demonstrated this effect in a clever map-task experiment with Mandarin L2-English speakers. All participants interacted with the same conversation partner (a research confederate) via audio, but half were told that the partner was a native speaker of American English, while the other half were told the partner was a non-native English speaker (with Swiss as L1). In reality the partner's speech was identical in both conditions. The results showed that the Mandarin speakers adjusted their pronunciation of certain vowels (/æ/ and /ɛ/) significantly more when they believed they were talking to a native speaker than when they thought their partner was another non-native speaker. They shifted their vowel qualities to more closely approximate native English norms when under the impression that their interlocutor was American. Simply believing one is talking to a native speaker triggered a stronger accommodation response in L2 production. The authors conclude that L2 speakers' production is not just passive imitation, but is modulated by social perception and attitudes – here, the prestige or status associated with a native speaker influenced how much the L2 learners adjusted their speech.

These studies on L2 production raise the question: How do such convergence tendencies relate to perception in L2 users? In bilinguals, the link between perception and production is complex. On one hand, accurate perception is generally thought to precede or enable accurate production—as one cannot reliably produce a sound distinction that one cannot hear. On the other hand, many late bilinguals can approximate native-like production of certain L2 sounds while still perceiving those contrasts through the filter of their L1 categories. This mismatch has been documented in classic work by (Caramazza et al., 1973) on French-English bilinguals production and perception of VOT. These bilinguals adjusted their VOT in production depending on whether they were speaking French or English, coming close to the monolingual norms in each language. However, in perception they did not fully switch to monolingual-like category boundaries. Instead, their perceptual identification of /pa/ vs. /ba/ sounds was intermediate, influenced by both their French and English experience. In short, they produced like two distinct monolinguals, but perceived like a bilingual, with a merged or dual set of criteria.

These observations motivate a key question for the present research: If bilingual listeners' productions are influenced by who they believe they are speaking to, might their perception likewise be affected by who they believe they are listening to? While prior work has documented perceptual mode-shifting in response to overt language context (Caramazza et al., 1973), this study investigates whether more subtle social cues, such as the perceived native language or ethnicity of the speaker, can also prompt shifts in bilinguals' phonetic boundaries. In doing so, the current research examines whether socially modulated accommodation, well established in production, extends into the domain of perception, and whether L2 listeners' phonetic category settings remain stable or adapt dynamically in response to speaker identity. These questions not only highlight the need to study perception directly, but also invite a closer examination of how bilinguals mentally represent and access their phonetic categories. If social context can modulate where a listener draws a boundary between two sounds, it may suggest the existence of multiple, language-specific systems that can be selectively engaged. The next section explores this possibility by reviewing theoretical accounts of bilingual phonological organization and category flexibility.

1.4 Bilingual Phonetic Architecture and the Double Boundary Hypothesis

Understanding how bilinguals perceive phonetic contrasts (like voicing) requires addressing a broader theoretical question: How are two phonological systems represented in the bilingual mind? Empirical findings suggest that bilinguals often maintain functionally distinct categories across languages, yet cross-linguistic influence is common. For example, (Caramazza et al., 1973) observed that bilinguals produced language-appropriate VOTs, suggesting separate systems in production, while their perceptual boundaries were intermediate, pointing to merged or interactive representations. These results indicate that bilingual phonetic systems may operate in a flexible and interconnected manner rather than as completely separate entities.

This flexibility is further captured in (Flege, 1995)'s Speech Learning Model (SLM). According to SLM, when an L2 sound closely resembles an L1 category, bilinguals may "equate" the two, leading to a merged perceptual category. If the sounds are sufficiently distinct, a new category can form. For instance, English /t/ may be similar enough to Mandarin /t/ (both aspirated) to be treated as equivalent, but differ enough from French /t/ (unaspirated) to merit a separate category. The model thus predicts asymmetries in L2 acquisition depending on whether a contrast maps onto an existing L1 category.

Applied to voicing, this reasoning suggests that Russian–English bilinguals may either equate English voiceless stops with their Russian counterparts (which have shorter VOTs) or form new categories. In the former case, a single broad category may span a wide VOT range, making it harder to detect distinctions like /b/ vs. /p/. With experience, bilinguals may split such categories or refine internal boundaries.

This flexibility—the ability to shift or recalibrate boundaries—has become a key construct in models of bilingual speech perception.

The Double Phonemic Boundary Hypothesis (Caramazza et al., 1973; Elman et al., 1977; Garcia-Sierra et al., 2009) proposes that bilinguals can adopt different category boundaries for the same phonetic continuum depending on context. (Elman et al., 1977) provided early evidence for this model by showing that Spanish-English bilinguals shifted their VOT boundary depending on whether the task was conducted in English or Spanish. Similarly, (Bohn and Flege, 1993) also found that Spanish–English bilinguals shifted vowel boundaries based on the task language. (Garcia-Sierra et al., 2009) further showed neurophysiological differences in bilinguals' responses to the same stimuli under different language contexts. Collectively, these studies suggest that bilinguals' perceptual systems can adaptively tune to one language or the other.

This dissertation tests whether social context cues can also trigger such tuning. While prior work has demonstrated mode-shifts via linguistic context (e.g., surrounding language or task instructions), the design in the present study examines whether these effects extend when social cues such as a talker's perceived L1 background, racial appearance, gender, and other identity markers are presented as speaker guises. For example, will a Mandarin–English bilingual categorize an ambiguous /ta/–/da/ token differently if they believe the speaker is American versus Chinese?

If bilinguals systematically adjust their phonemic boundary based on speaker identity, it would extend models like the double-boundary hypothesis to include socially mediated switching. It would also suggest that bilinguals' multiple phonetic systems can be selectively engaged not only by explicit cues, but also by implicit social beliefs. Conversely, if no perceptual shift is observed, it would imply that for this contrast and group, listeners rely on a stable category—perhaps a compromise between L1 and L2 norms. Either outcome offers insight into how bilingual phonetic categories are structured and whether they remain fixed or adapt dynamically under social influence.

1.5 Research Questions and Hypotheses

Building on the theoretical and empirical foundations outlined above, this study investigates how Mandarin-English and Russian-English bilingual listeners integrate social and acoustic information when processing sounds along a VOT continuum. Specifically, it examines whether listeners' beliefs about a talker's native language background modulate their categorization of English voicing contrasts under various social conditions. To address this overarching question, the dissertation evaluates three dimensions of perceptual behavior:

1. **Explicit boundary judgment (Chapter 3).** How do bilingual listeners rate voicing contrasts in isolated syllables when the supposed native-language status of the speaker is manipulated? Using a Visual Analog Scale (VAS) task, this experiment tests whether labeling a talker as a native, familiar non-native, or unfamiliar non-native English speaker shifts where listeners "hear" the category boundary between /ba/-/pa/, /da/-/ta/, and /ga/-/ka/.

2. **Implicit processing in context (Chapter 4).** How do bilingual listeners process voicing contrasts in real words when primed with audiovisual cues about the speaker's ethnicity and language background? Using anticipatory eye movements during real-time lexical processing (e.g., BARK vs. PARK), this experiment examines whether listeners' gaze behavior reflects socially modulated categorization.

3. **Individual differences (Chapter 5).** To what extent do listener-specific traits—such as English oral proficiency or autistic-like cognitive style—influence the degree to which social cues modulate phonetic perception?

These questions are tested across two groups of late sequential bilinguals: Russian-English and Mandarin-English speakers. These groups were selected because their L1s represent opposing ends of the VOT continuum: Russian has relatively short-lag voiceless stops, while Mandarin has long-lag voiceless stops. Beyond these phonetic contrasts, the groups also differ in their prototypical racial representation: Mandarin speakers are typically perceived as nonwhite, where as Russian speakers are often perceived and self-categorized as white. This distinction matters because race is a salient component of social guises (Kutlu et al., 2022c,b; Kutlu, 2023). Thus, the selection of these two speaker groups not only enables testing perceptual boundary shifts from both ends of the VOT spectrum, but also allows examination of how racialized guises interact with bilingual listeners' perception of voicing contrasts, a theme explored further in in Chapters 3 and 4. As a result, any shift in category boundary due to social information should manifest in opposite directions between the two groups, offering a strong test of the effect.

Across all experiments, participants were presented with auditory stimuli derived from English voicing minimal pairs across three places of articulation (bilabial, alveolar, and velar), each synthesized into a nine-step VOT continuum spanning from clearly voiced to clearly voiceless. After listeners completed baseline judgments on auditory stimuli alone, they then responded to the same stimuli paired with audiovisual guises representing three speaker identities: American, Chinese, and Russian.

This study hypothesizes that bilingual listeners' perceived phonetic boundaries in English are not fixed, but shift in predictable ways depending on both the social context and the listener's own language background. Specifically, listeners' categorical boundaries between voiced and voiceless sounds are expected to lie intermediate between English and L1 norms, reflecting bilinguals' mixed phonetic experience. Moreover, different talker guises are predicted to modulate which phonological system activates. For instance, when listening to a speaker believed to be a fellow L2 English user (e.g., with Russian or Chinese as L1), listeners may draw on their L1-tuned boundary (e.g., shifting toward shorter VOT for Russian listeners, longer VOT for Mandarin listeners). In contrast, a native-speaker guise may cue a more English-like categorization. In essence, social identity cues may tilt bilingual perception toward one or the other of their two phonological systems. Observing such shifts would provide strong evidence for the double phonemic boundary hypothesis in a socially cued context; failure to observe such shifts would support a single-system interpretation of bilingual phonetic perception.

Having introduced the aims of this research, the next chapter (Chapter 2) provides a targeted literature review grounding the experimental design. It covers foundational work on the matched guise task, crosslinguistic differences in voice onset time (VOT), individual difference measures known to explain within-group listener variability, and the rationale for using eye-tracking to capture real-time processing. This context motivates the three experiments and clarifies their place in the broader landscape of bilingual speech research.

The remaining chapters report the three empirical studies and their implications. Chapter 3 tests whether talker identity modulates listeners' explicit phonetic judgments in a VAS rating task. Chapter 4 examines whether online perceptual processing is likewise guided by social expectations, using anticipatory eye movements. Chapter 5 explores how listener traits shape perceptual plasticity in the face of social cues.

Finally, Chapter 6 summarizes the results across tasks and discusses broader implications for theories of bilingual phonetic perception and socially driven variability.

Chapter 2

LITERATURE REVIEW ON METHODOLOGICAL INTEGRATIONS

The chapter reviews literature that highlights four critical components. First, matched-guise paradigms (Section 2.1) provide a controlled way to probe how beliefs about speaker identity shape perception, particularly in ambiguous contexts. Second, voice onset time (VOT; Section 2.2) is a robust and cross-linguistically variable cue, making it ideal for testing bilingual listeners' perceptual boundaries and how these boundaries shift under social modulation. Third, individual differences (Section 2.3) remind us that bilinguals are not a homogeneous group: proficiency and cognitive style, particularly autism-aligning traits, may explain why social information influences some listeners more than others. Finally, eye-tracking (Section 2.4) offers a dynamic, implicit measure of perceptual processing, capturing the time course of social effects in ways that explicit rating tasks cannot.

2.1 Matched-Guise Paradigms as Assessment for Social Modulated Speech Perception

The matched-guise paradigm has long served as a powerful methodological tool for uncovering both how listeners' beliefs about a speaker shape their perceptual responses to speech and how their form social evaluations of the speaker based on the characteristics of the speech (Campbell-Kibler, 2007; Marcinko, 2023). This paradigm manipulates listeners' expectations about speaker identity while holding the speech signal constant, thus isolating the top-down effects of social information. In its original form (Lambert et al., 1960), it was used to probe attitudinal judgments by having bilingual speakers produce the same content in two languages, revealing that listeners attributed different personality traits depending on the language, interpreting language as a proxy for speaker identity. Since then, matched-guise methods have been widely adapted to examine more subtle perceptual phenomena, including accent evaluation (Nejjari et al., 2019), speech style (Tamminga, 2017), and phonetic categorization (Reinisch and Holt, 2014).

Matched-guise tasks are particularly well suited to studying perceptual boundary shifts—cases where acoustic cues are ambiguous and subject to top-down influence. For example, when listeners are asked to categorize fricatives along an /s/-/ʃ/ continuum, knowing the presumed gender of the speaker can bias judgments — with the same ambiguous sound perceived as as postalveolar /ʃ/ when paired with male speaker and /s/ with female speaker, consistent with socially indexed speech norms (Strand, 1999). Similarly, (Hay and Drager, 2010) showed that vowel categorization could shift when the same speech was attributed to speakers from different dialect backgrounds. These effects are especially likely to emerge in mid-range steps of continua, where acoustic cues are most ambiguous and prior expectations have the most room to influence perception.

As detailed in Chapter 1 and above, listeners often show different perceptual responses to the same speech signal depending on whether the talker is introduced via a photo, a written description, or other cues, e.g., (Hay and Drager, 2010; Johnson et al., 1999; Niedzielski, 1999; Rubin, 1992). These studies confirm that listeners' social expectations can recalibrate the interpretation of ambiguous sounds. The present section builds on this conceptual foundation and focuses on how these methods have been leveraged to design experiments that probe phonetic categorization, especially in contexts of ambiguity.

The present study adopts this approach to test how audiovisual social guises highlighting talkers' native language background modulate L2 listeners' phonetic boundary placement in categorizing stop consonants by VOT. As previously discussed in Chapter 1, socially induced bias in bilingual perception may arise either from expectations about talker-specific phonetic norms or from listeners' own metalinguistic assumptions about who "should" produce which sounds. By using identical audio stimuli across guises, this design can isolate how beliefs about speaker identity influence perceptual decisions.

This approach also enables several analytical advantages. First, it allows us to focus on the ambiguous regions of the VOT continuum, where bottom-up input is insufficient to guarantee a stable categorization and where social information may tip the balance. Second, it permits an examination of asymmetries in bias—for instance, whether the native guise functions as a perceptual default, or whether the non-native guise elicits stronger modulation due to heightened listener expectations. Third, it opens the door to probing individual variability in susceptibility to social cues, which is explored in later chapters.

Finally, although attitudinal judgments are often intertwined with speech perception, the present study limits its focus to phonetic categorization, not overt attitudes. The matched-guise paradigm provides a subtle but controlled method for introducing social context without directly prompting evaluative responses. In doing so, it lays the groundwork for investigating how bilingual listeners resolve ambiguity in speech when faced with minimal but salient social cues.

2.2 Voice Onset Time as a Phonetic Cue and Experimental Probe

Voice Onset Time (VOT) is the interval between the release of a stop consonant and the onset of vocal fold vibration (voicing) for the following vowel. In many languages, it serves as a fundamental cue for differentiating voiced and voiceless stop consonants. For example, in English, the sounds /b, d, g/ (voiced stops) typically have a short VOT (often even a slight voicing lead or near-zero lag), whereas /p, t, k/ (voiceless stops) have a longer VOT due to a burst of aspiration. The boundary at which a listener stops hearing a stop as voiced and starts hearing it as voiceless is a classic measure in phonetic perception. VOT is well-suited as an "experimental probe" for several reasons:

- **Acoustic measurability.** Acoustic measurability: VOT is straightforward to manipulate and quantify in a controlled way. One can synthesize or digitally alter recordings to create continua of VOT durations spanning from very short (fully voiced) to very long (long-lag aspirated). This quality allows precise control over stimulus ambiguity and ensures the ability to finely sample the perceptual crossover region.

- **Perceptual salience.** The voicing contrast (e.g., hearing "da" vs. "ta") is perceptually salient and familiar to listeners, meaning participants can reliably perform identification tasks on these sounds. Yet by adjusting VOT, the distinction can be made as easy or as hard as needed (facilitate the piloting process). The unambiguous tokens (clearly voiced or clearly voiceless) are included to ensure listeners understand the task (clearly voiced or clearly voiceless syllables), and ambiguous ones are used to test the subtle effects.

- **Cross-linguistic variability.** Crucially for the present study, different languages implement voicing contrasts with different typical VOT values. English is often characterized as having a "long-lag" voiceless set (roughly $\sim 50 - 80$ ms VOT for /p, t, k/) and short-lag (near $0 - 20$ ms) for /b, d, g/. Russian, by contrast, has a short-lag voiceless system (around $15 - 30$ ms for /p, t, k/) and often pre-voicing (negative VOT) for /b, d, g/. Mandarin Chinese also has long-lag aspirated vs. short-lag unaspirated stops, but importantly both categories in Mandarin are phonetically voiceless (Mandarin does not use actual voicing in this contrast). In practical terms, Mandarin /p, t, k/ have very long VOTs (often $60 - 90$ ms), while Mandarin /b, d, g/ (voiceless unaspirated stops) have VOTs around $10 - 20$ ms. These cross-language differences mean that a given intermediate VOT might straddle category boundaries differently for different listeners. Table 2.1 summarizes typical VOT ranges in Russian, English, and Mandarin, based on values reported in the literature.

Table 2.1: Summary of VOT ranges for voiceless aspirated stops in Russian, English, and Mandarin.

Russian (ms)			English (ms)			Mandarin (ms)		
(Ringen and Kulikov, 2012)			(Lisker and Abramson, 1964)			(Chao and Chen, 2008)		
/p/	/t/	/k/	/p/	/t/	/k/	/p/	/t/	/k/
18	20	38	58	70	80	82	81	92

The above properties make VOT an ideal test case for bilingual speech perception. It is a single, easily manipulable acoustic dimension whose linguistic interpretation varies across languages. This quality allows the present study to examine whether bilinguals shift their perceptual boundary depending on context. I predict that listeners may effectively maintain two reference points—one aligned with their L1 and another with English—and that social cues may dynamically bias perception toward one or the other. For instance, a Russian–English bilingual may more readily categorize short-lag VOTs as voiceless when the talker is believed to be Russian, but shift toward a longer VOT threshold under a native English guise. Mandarin–English bilinguals may show the reverse pattern, aligning with longer VOT expectations under a Mandarin guise. In short, bilinguals' voicing boundaries are expected to be context-sensitive when social information cues one language over another.

Another aspect that makes VOT a useful probe is that it also lends itself to testing perceptual asymmetries. English listeners are generally more sensitive to increased aspiration (longer VOTs) than to its reduction, often failing to detect shortened VOTs (Nielsen and Scarborough, 2015), and they tend to imitate lengthened VOTs more than shortened ones in production (Nielsen, 2011). Whether this asymmetry is universal or language-specific remains debated, but in a bilingual context, it raises new questions: Does the same asymmetry hold for L2 listeners? Can social expectations modulate sensitivity? For example, expecting a non-native talker might heighten sensitivity to shorter VOTs, reinforcing accent expectations, or it might dampen sensitivity by normalizing the variation. VOT thus provides a controlled means of empirically testing these possibilities.

2.3 Listener Background and Individual Differences

While the previous section focused on group-level patterns rooted in cross-linguistic phonetic experience in VOT norms between Mandarin, Russian, and English, such patterns tell only part of the story. Bilinguals are a highly heterogeneous population and within any bilingual population, listeners do not behave uniformly. They differ in language proficiency, usage history, and social attitudes, all of which may modulate how they respond to ambiguous stimuli or social cues. Recent research in sociolinguistics and speech science has emphasized that group-level averages can mask substantial within-group variability, shaped by experience, cognitive style, and individual dispositions. In this spirit, the present study moves beyond a simple Mandarin vs. Russian bilingual comparison and asks: which individuals, regardless of group, show stronger or weaker effects of social information?

This question is important not only methodologically, but ethically. Overgeneralizing from group means can lead to essentialist or even harmful assumptions, such as attributing perceptual tendencies solely to ethnicity or L1 background. Third-wave sociolinguistics reminds us that individuals are active agents, not fixed representatives of a social category. Within each L1 group, we expect a range of behaviors, possibly shaped by factors like English proficiency, daily usage, or cognitive style. By identifying such variation, this dissertation aims to develop a model of socially modulated bilingual speech perception that reflects listeners' lived reality.

2.3.1 Language Experience and Proficiency

One major factor shaping perception is L2 experience and proficiency. More balanced or proficient bilinguals may show greater flexibility in switching phonemic boundaries between languages. (Elman et al., 1977) found that the magnitude of phonetic boundary shifts correlated with degree of bilingualism. A highly proficient bilingual is likely to have more robust L2 phonetic categories, potentially reducing reliance on social cues—or alternatively, they may be more adept at using context when needed. (Roberts, 2012) reviewed studies showing that highly proficient L2 learners typically process language in a native-like way unless the task requires metalinguistic awareness, at which point individual differences like working memory or proficiency become more apparent. Translating this consideration to our study: in an explicit task like the VOT continuum rating (Chapter 3), listeners with higher English proficiency may set more native-like category boundaries and may integrate social cues differently than less proficient listeners (e.g., more reliably and accurate activate L2 English phonology when presented with an American guise).

To explore these effects, participants' English oral proficiency is assessed using the English Elicited Imitation Task (EIT), a validated shortcut oral proficiency test that demonstrates strong alignment with L2 learners oral proficiency results from more elaborate standardized assessments (Kostromitina and Plonsky, 2022; Tracy-Ventura et al., 2014; Wu and Ortega, 2013), thereby providing a robust, objective approximation of their productive command of English.

2.3.2 Cognitive Styles and the Role of Autism-aligning Traits on Speech Perception

Cognitive factors are another likely source of variation. Abilities such as working memory, attentional control, and inhibition have been linked to how well listeners adapt to difficult or ambiguous speech. For example, Ou and colleagues (Ou et al., 2015; Ou and Law, 2017) found that Cantonese-speaking listeners with stronger executive control were better at distinguishing lexical tones and maintaining category boundaries. (Lev-Ari and Peperkamp, 2013) similarly found that bilinguals with weaker inhibitory control showed more cross-language interference; French–English bilinguals with poor inhibition produced and perceived English with more French-like features, likely due to an inability to suppress their L1.

A particularly intriguing factor for us is cognitive style as reflected in variation across social and attentional domains. The Autism Spectrum Quotient (AQ) is a self-report questionnaire designed to assess the degree to which adults with normal intelligence exhibit social and cognitive tendencies associated with autism-specturm-related traits (Lundqvist and Lindner, 2017). The instrument comprises 50 items covering five sociocognitive subscales — social skills, attention switching, attention to detail, communication, and imagination — providing a quantitative index of individual differences in cognitive processing and interaction style. In this dissertation, the AQ is used not as a diagnostic instrument but as a tool for examining how these subdomains of sociocognitive styles may shape bilinguals' speech perception, capturing variation beyond listeners' language background and English proficiency.

In recent decades, the use of the AQ in speech perception research has revealed systematic links between autism-aligned traits and how listeners adapt to speech variability. For example, (Yu, 2010) found that women with lower AQ scores (i.e., those who socially and cognitively diverge from autistic populations) actually exhibited less perceptual compensation for a coarticulatory cue, compared to men and higher AQ individuals. In that study, listeners with very low autistic-trait levels failed to discount a contextual vowel effect that they normally "should" have (a process akin to accounting for coarticulation), whereas those with higher autistm-aligned traits showed more context-normalization. This counterintuitive finding suggests that having a more detail-focused (and perhaps less socially driven) cognitive style can sometimes enhance certain perceptual adjustments. Other work by (Yu et al., 2011) similarly showed that autism-aligned traits and working memory capacity together affected how people used phonotactic cues in perception. Meanwhile, (Stewart and Ota, 2008) reported that individuals with higher autism-aligned trait levels were less influenced by lexical knowledge when identifying speech sounds, sticking more closely to acoustic input rather than top-down expectations. These studies indicate that listeners who are more systemizing or detail-oriented (traits linked to the autism spectrum) may rely less on higher-level contextual or social information when perceiving speech. Conversely, those who are more intuitive or socially oriented might be more prone to let context sway their perception, possibly at the expense of fine acoustic detail.

Building on these findings, this dissertation incorporates the Autism Spectrum Quotient (AQ) for Adults to examine how sociocognitive style modulates socially driven shifts in bilingual speech perception. By assessing each participant across five subscales (Social Skills, Attention Switching, Attention to Detail, Communication, and Imagination), I aim to determine whether individuals with higher AQ scores show reduced susceptibility to talker-based biases compared to those with lower AQ scores.

For reasons of scope and focus, detailed analyses of language attitudes, sociocultural identification, and social networks are not included in this dissertation. However, incorporating English oral proficiency and AQ measures will still enrich the interpretation of group-level results and help frame future work on socially modulated speech perception. The next section (2.4) discusses eye-tracking as a key methodological tool for speech perception research, highlighting its advantages for capturing moment-to-moment perceptual decisions and subtle biases that traditional tasks may overlook.

2.4 Eye-Tracking as a Tool for Speech Perception Research

Eye-tracking is a widely used method in second language acquisition (SLA), bilingualism, and psycholinguistics, offering fine-grained, real-time insights into how listeners process speech. However, its application to sociophonetic perception remains underexplored. This section reviews key advantages demonstrated by past studies, considers why eye-tracking has been underutilized in socially modulated speech perception (e.g., its costliness and technical demands compared to more accessible tasks like surveys, rating scales, or interviews), and outlines a new, cost-effective webcam-based approach to study bilingual listeners' perceptual biases.

Eye-tracking captures real-time, unconscious responses during language processing. By recording where and when listeners look while hearing speech, researchers can track how they incrementally interpret the acoustic signal before forming a conscious judgment or overt response. Because eye movements are largely involuntary, eye-tracking is less susceptible to social desirability or self-censorship, making it especially valuable for detecting implicit perceptual biases—the subtle ways that expectations and social context shape interpretation beneath participants' awareness. Traditional tasks like identification or rating capture only the final decision, whereas eye-tracking reveals the processing pathway leading to that decision.

A core method that exemplifies this power is the Visual World Paradigm (VWP). In a typical VWP experiment, participants view several pictures while hearing a spoken word, and their gaze shifts toward potential referents as the word unfolds. For example, hearing "cand-" might initially trigger looks to both candy and candle, until the final sound disambiguates the word. Classic work by (Tanenhaus et al., 1995) demonstrated that these gaze shifts provide a moment-to-moment record of comprehension, revealing syntactic and semantic processing well before participants make a conscious selection. This continuous, implicit measure has made eye-tracking a cornerstone of psycholinguistics and bilingualism research.

For bilinguals, eye-tracking has uncovered evidence of language co-activation. (Marian and Spivey, 2003) showed that Russian–English bilinguals, when hearing "marker," often looked at a stamp (marka in Russian), indicating parallel activation of both lexicons. Similarly, studies of code-switching have demonstrated that switching between languages mid-sentence incurs a processing cost, observable as delayed or disrupted gaze patterns, especially among bilinguals with less experience in code-switching environments (Valdés Kroff et al., 2018). These findings underscore eye-tracking's ability to reveal cross-language dynamics that traditional end-point tasks might overlook.

Despite these successes, eye-tracking remains underutilized in sociophonetic perception research. While rating tasks, surveys, or interviews are comparatively easy to deploy, eye-tracking has historically been costly and logistically challenging, requiring specialized equipment and controlled lab environments. Only two published sociophonetic studies have used eye-tracking – (D'Onofrio, 2018) and (Koops et al., 2008) – both of which revealed perceptual biases that traditional measures failed to capture. (D'Onofrio, 2018) showed that gaze patterns uncovered subtle associations between TRAP-backing and Californian identity that were absent in matched-guise evaluations, while (Koops et al., 2008) found that gaze data revealed implicit differences in PIN/PEN merger perception linked to talker age, even when explicit judgments showed little variation. These results suggest that eye movements can reveal early, implicit activation of sociolinguistic knowledge, effects that might otherwise remain hidden.

In this dissertation, Experiment 2 employs a webcam-based anticipatory eye movement (AEM) paradigm to extend eye-tracking into the study of bilingual speech perception with social cues. Participants learn a simple left/right visual association with voiced vs. voiceless categories (e.g., /b/-initial words on the left, /p/-initial words on the right). On critical trials, they hear ambiguous VOT stimuli (e.g., between beach and peach), and their first gaze shift indicates the category they implicitly perceive—before any conscious response is made. By pairing these stimuli with speaker guises (native vs. non-native), the design tests whether social information biases this early perceptual resolution.

To make eye-tracking accessible beyond the lab, this study uses OpenFace 2.0 (Baltrušaitis et al., 2016), an open-source gaze estimation toolkit that performs facial landmark detection and gaze tracking from ordinary webcam recordings. This approach significantly lowers costs and logistical barriers while maintaining sufficient accuracy for left/right gaze tracking as employed by the experiments in this dissertation. By integrating this tool, the study not only investigates the interaction of social and acoustic cues but also demonstrates the feasibility of scalable, low-cost eye-tracking for sociophonetic research.

In summary, this eye-tracking paradigm complements the explicit Visual Analog Scale task (Chapter 3) by providing time-sensitive, implicit data about how bilinguals integrate social expectations with acoustic cues. Chapter 4 details the implementation of this method and examines whether talker identity affects perception at the earliest stages of phonetic processing, capturing effects that might otherwise remain hidden.

2.5 Summary of Literature and Research Plan

The preceding sections have outlined the theoretical and methodological foundations that motivate this dissertation. Chapter 1 introduced the central research questions, emphasizing how bilinguals' phonetic boundaries may shift as a function of both cross-linguistic phonetic experience and social information about the talker. Chapter 2 built on this foundation by reviewing key methodological tools and experimental paradigms suited to investigating these questions. Together, these insights inform the experimental design of this dissertation. The next chapters translate the conceptual and methodological groundwork of Chapters 1 and 2 into empirical tests of how bilingual listeners integrate social expectations and acoustic cues during speech perception.

Chapter 3

NAVIGATING SOCIALLY MODULATED VOICING CONTRAST PERCEPTION IN PRE-LEXICAL CONTEXT: A VAS STUDY

3.1 Introduction

Speech perception is a highly dynamic process shaped by the interplay between bottom-up acoustic signals and top-down expectations, e.g., (Zekveld et al., 2006; Davis and Johnsrude, 2007; Ohala, 2014; Wade, 2022). Listeners do not passively receive sounds; rather, they actively use linguistic and social knowledge to impose structure on the sensory input they encounter (Niedzielski, 1999). A growing body of research demonstrates that listeners' beliefs about a talker's background—such as their perceived gender (Strand, 1999; Ceuleers et al., 2022), age (Gordon et al., 2019), regional (Campbell-Kibler, 2007; Hay and Drager, 2010), ethnic (Kutlu et al., 2022c,b), or linguistic identity—can influence how incoming speech sounds are categorized, sometimes altering perception even when the acoustic signal remains unchanged (Lambert et al., 1960). However, much of the existing work on social speech perception has focused on L1 listeners processing contextualized speech (Campbell-Kibler, 2007; Hay and Drager, 2010), such as full sentences or naturalistic conversations. Less is known about how L2 bilingual listeners integrate social information during perception of isolated acoustic features (e.g., VOT continua), where no broader sentence context is available to scaffold interpretation. Critically, examining how top-down social expectations warp perception of acoustically minimal pairs can reveal whether social priors operate at pre-lexical processing stages, penetrating foundational phonetic encoding before word recognition occurs.

The present experiment addresses this gap by investigating how bilingual listeners categorize fine-grained phonetic contrasts — specifically, the voicing distinction between three stop sound pairs: /ba/ and /pa/, /da/ and /ta/, /ga/ and /ka/ — when exposed to social cues about the talker's language background. By embedding stimuli in syllable-only contexts, the experiment isolates low-level acoustic processing from higher-order semantic or syntactic influences, allowing a direct examination of how top-down social expectations interact with basic phonetic categorization.

The Visual Analog Scale (VAS) method was chosen because it captures participants' gradient categorization decisions along a continuous scale, rather than forcing discrete choices. This approach is especially valuable for detecting subtle shifts in perception that may arise from the interplay/interaction between social information and acoustic signals like VOT, unlike 2-alternative forced-choice tasks that mask hesitation (Kutlu et al., 2022a). By combining social priming (through guise manipulation of the talker's supposed L1 background) with acoustic variation (a VOT continuum from /ba/ to /pa/), this experiment aims to uncover how bilingual listeners' phonological activation is influenced by social expectations, and whether social cues trigger L1- versus L2-based perceptual patterns even at the level of isolated syllables.

This chapter systematically details the experimental trajectory, beginning with stimulus design and social guise implementation, progressing through analytical methodologies, and culminating in empirical findings. I demonstrate how Mandarin-English and Russian-English bilinguals dynamically recalibrate voicing boundaries when social identity modulates acoustic processing, revealing that perceptual reorganization occurs not through broad biases but through acoustically constrained mechanisms. Importantly, these effects emerge even for isolated syllables, establishing social meaning as a fundamental architect of pre-lexical phonetic representation.

3.1.1 Research Questions and Hypotheses

Our experiment addresses the following questions about bilinguals' L2 English listeners' processing of voicing contrasts in syllable context:

1. Do Russian-English and Mandarin-English bilinguals' voicing contrast boundaries in L2 English differ from the VOT boundaries reported for their L1s in previous literature? This question is addressed using results from the *baseline* (no social cues) condition.
   (a) $\boldsymbol{H_0}$(*null*): No, bilinguals will maintain L1-consistent perceptual boundaries when processing syllables in L2 English.
   (b) $\boldsymbol{H_1}$ (*alt*): Yes, bilinguals will show evidence of cross-linguistic influence
2. Do bilinguals' categorization of voicing contrasts vary as a function of **both** VOT duration and perceived talker identity (social guise)? This question investigates whether bilingual listeners rely solely on bottom-up acoustic cues or also incorporate top-down social information.
   (a) $\boldsymbol{H_0}$: Listeners' judgments will depend only on acoustic VOT values, with no effect of talker guise.
   (b) $\boldsymbol{H_1}$: Talker guise will modulate perception, such that listeners' voicing judgments are influenced by any given social information about the speaker's language background.
3. If social cues do affect voicing perception, would these shifts systematically align with listeners' L1 and L2 phonological expectations?

(a) $\boldsymbol{H_0}$: All social effects will be random or inconsistent in direction.

(b) $\boldsymbol{H_1}$: Social cues will shift perceptual boundaries in predictable directions. For example, Russian-English bilinguals will show earlier VOT boundary shifts (closer to Russian norms) when presented with a Russian speaker guise, and later VOT boundary shifts (closer to English norms) under an American guise.

3.2 Methods

3.2.1 Overview

This experiment assesses listeners' explicit judgments of three stop consonant pairs: /pa/-/ba/, /ta/-/da/, and /ka/-/ga/. To elicit these judgments, I use a modified Visual Analogue Scale (VAS) paired with socially cued speaker guises. The guises reflect three language backgrounds: (1) a native speaker of Mandarin Chinese, (2) a native speaker of American English, and (3) a native speaker of Russian. Thus, each listener group is exposed to three social guises: a *familiar non-native* guise matching their L1 (e.g., Mandarin for Mandarin-English listeners), a *native* English guise (L2), and an *unfamiliar non-native* guise mismatched with their L1 (e.g., Russian for Mandarin-English listeners).

This unfamiliar non-native guise serves as a critical control condition. Its inclusion allows us to test whether perceptual shifts are uniquely triggered by socially familiar cues that align with listeners' L1 phonology, or whether the mere presence of any non-native identity, regardless of familiarity, can influence perception. If perceptual boundaries shift only in response to familiar guises, this would suggest that socially driven phonological activation is selective and systematically aligned with listeners' internalized phonological categories. In contrast, if the unfamiliar non-native guise also induces perceptual change, this may indicate a broader susceptibility to social information or a less specific form of top-down cue integration. Moreover, if listeners accurately adjust their perceptual boundaries in response to the *unfamiliar non-native* guise (e.g., Chinese listeners categorizing an item with shorter VOT under Russian condition than they do under both the Chinese and the American condition), then this would reveal that bilinguals can rapidly form novel social-phonological mappings from brief exposure and apply these adapted boundaries even in semantically minimal, decontextualized syllable processing.

Speaker guise is operationalized through audio-visual input: participants first watch a series of short videos featuring speakers with the target backgrounds, and are then told that the talker they are about to hear is from the same language background as those in the video clips. All participants complete a baseline condition and all three social conditions (guises), using a within-subject design. The experiment consists of four phases:

1. **Training phase.** Participants practice rating visual and non-target auditory stimuli on the VAS.

2. **Baseline phase.** Participants rate the same auditory stimuli without any social cues.

3. **Guise familiarization.** Participants are introduced to three speaker guises via video.

4. **Experimental phase.** Participants rate the same auditory stimuli as in the baseline, now accompanied by speaker guise cues.

Order of the social guises is randomized per participant, and within each guise, all 81 audio trials are completed before proceeding to the next. Stimuli are not intermixed between guises. Each unique auditory stimulus is presented three times, across a 9-step VOT continuum and three places of articulation, resulting in 81 trials per condition. Across the baseline and three social conditions, participants complete 324 trials in total. The order of VOT steps within each condition was fully randomized, and the sequence of social guise blocks and the order of places of articulation of the target continua were also randomized across participants. I built and administered the experiment using PCIbex (Zehr and Schwarz, 2018), a browser-based experiment platform built on JavaScript syntax. See Figure 3.1 for an illustration of the experiment flow.

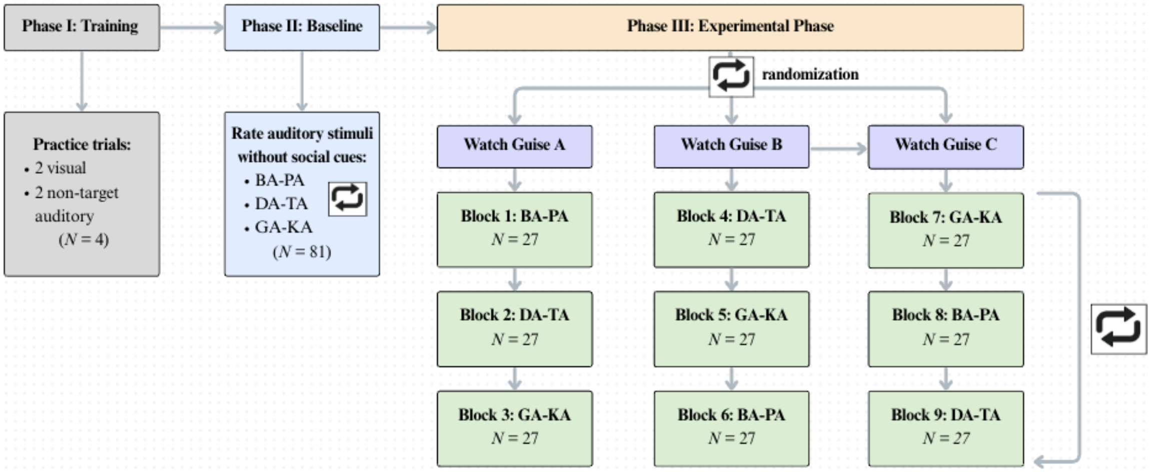


Figure 3.1: Flowchart showing the structure of one version in the visual analogue scale (VAS) task.

The three phases include: training, baseline, and three interleaved guise familiarization followed by rating blocks. Each guise (e.g., A = Chinese, B = American, C = Russian) precedes auditory rating blocks for three POAs (BA–PA, DA–TA, GA–KA), with guise assignment and POA order pseudo-randomized across three versions.

3.2.2 Materials

Creation of speaker guises

Impressions of speaker guises were constructed using a series of short video clips publicly available on YouTube. Videos were selected to feature real people speaking naturally in English with their own regional or L1-accented speech. Each guise includes 3–5 short clips, and speakers represent the intended language background through setting, topic, and accent.

- **American guise**
    - Buzzfeed video featuring a diverse group of Americans discussing sandwiches.
    - Street interviews about travel and opinions on America.
- **Chinese guise**
    - Street interviews conducted in Shanghai, themed around learning English and studying abroad.
- **Russian guise**
    - Vlogs and interviews filmed in Moscow, discussing Russian culture, travel, and language learning.

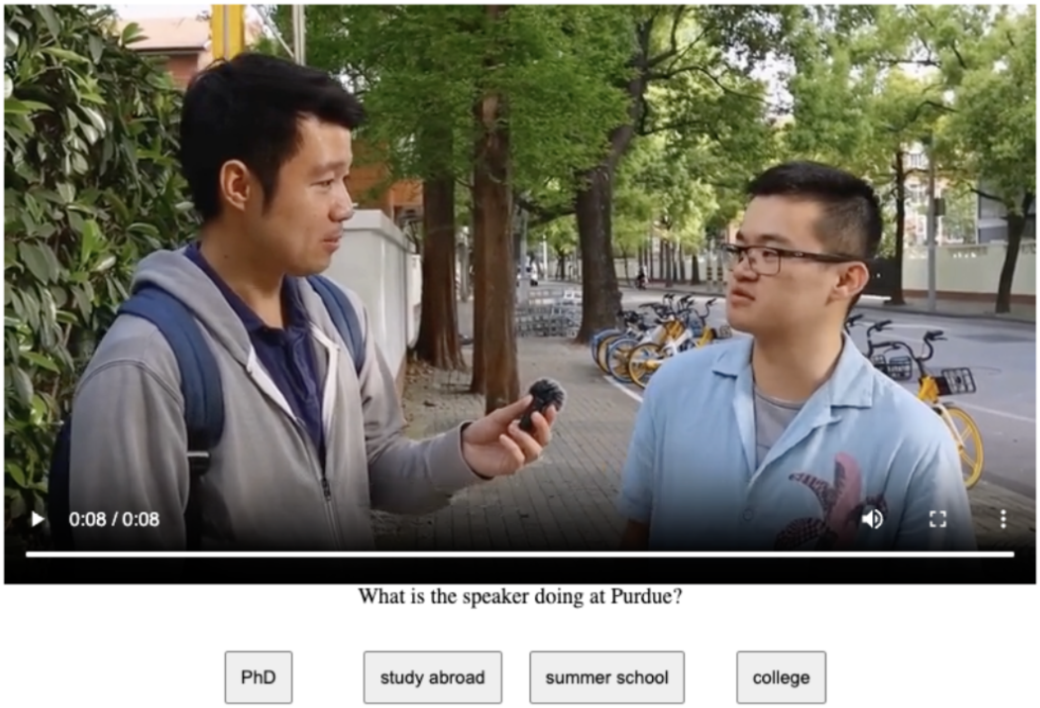
I carefully excluded any videos in which speakers appeared to be exaggerating, imitating, or performing an accent, as such portrayals risk reinforcing linguistic stereotypes and compromise the integrity of sociophonetic research. Our goal was to present naturalistic, authentic speech that reflects the genuine linguistic background of each speaker. To this end, I selected videos that were balanced across speaker guises in terms of thematic content (e.g., discussions of culture, language, travel, and food), tone, and visual quality. Additionally, I ensured that stop-initial tokens were comparably distributed across guises to avoid skewing perceptual input. Table 3.1 summarizes the background of each selected speaker and the number of stop-initialed tokens produced by each speaker and total for each guise.

Table 3.1: Summary of guise videos: speaker L1, gender, video length, and number of stop-initial tokens.

Guise	Video File Name	Gender	Length (s)	Stop-Initial Tokens
Mandarin				
	`CN-1-female`	Female	12	1
	`CN-2-male`	Male	12	2
	`CN-3-male`	Male	14	0
	`CN-4-male`	Male	14	4
	`CN-5-male`	Male	8	2
	Total	*5 speakers*	*60*	*9*
Russian				
	`RU-1-male`	Male	9	4
	`RU-2-male`	Male	19	6
	`RU-3-male`	Male	6	0
	`RU-4-male`	Male	14.7	1
	`RU-5-female`	Female	14	2
	Total	*5 speakers*	*62.7*	*13*
American				
	`AM-1-male`	Male	9	3
	`AM-2-female`	Female	13	2
	`AM-3F-4M-5F-6M`*	2 Female, 2 Male	26	16
	Total	*6 speakers*	*48*	*21*

* *AM-3F-4M-5F-6M comprises* 2 *female and* 2 *male speakers.*

To maintain participants' engagement and ensure attentiveness during the guise familiarization phase, a multiple-choice comprehension question occasionally followed a guise video. These questions were presented unpredictably, rather than after every video, to keep participants alert and discourage passive viewing. In addition to verifying comprehension, this unpredictability helped sustain participant focus and introduced brief delays between the guise familiarization and auditory rating blocks, thereby reducing the likelihood that participants would recognize they were hearing the same voice across all conditions. See Figure 3.2 below for an example from the Mandarin guise phase.



What is the speaker doing at Purdue?

Figure 3.2: Mandarin phase guise familiarization. A screenshot from the Mandarin guise familiarization phase, showing the video presentation followed by a multiple-choice comprehension and attention check question.

Creation of Auditory Stimuli

This section details the development of auditory stimuli, from initial recordings and pilot testing to the techniques used for creating the synthesized auditory continuum.

Initial Recording. Four native speakers were invited to record the six target syllables (/pa/, /ba/, /ta/, /da/, /ka/, /ga/) in the sound booth in the Linguistics Lab at Georgetown University. All four speakers are American English speakers from different U.S. regions (Georgia, North Carolina, Seattle, and Hawaii) and are advanced bilinguals or multilinguals in at least one other language. Recording was obtained with Zoom h4n Handy Pro recorder in wav format to preserve acoustic details. Syllables were presented to them in black text against white background, one word per slide. The order of syllable presentation was randomized across speakers

Selecting the Neutral Voice. To ensure that perceived differences could be attributed to social information alone rather than acoustic differences between talkers, I selected one talker who was judged to be maximally place-neutral. To evaluate this perceived social identity, I invited 21 raters to judge short mono-syllabic tokens from each speaker, rating how likely each speaker sounded Russian, Chinese, or American on a $0 - 10$ scale. Raters included 11 native Mandarin speakers and 8 native Russian speakers, with the rest highly familiar with both Mandarin and Russian speakers via personal or professional contexts.

I calculated both the mean and median of each speaker's ratings across all three guises. Then, I selected the speaker whose ratings were consistently closest to the neutral midpoint (i.e., 5 on a $0 - 10$ scale) across all conditions as the model talker. This dual-metric approach helped avoid bias from skewed judgments and ensured that the selected voice was least marked for any specific social identity. Our winning auditory stimuli came from Michael, a native speaker of American English from North Carolina and a graduate student in the Spanish Department.

Creation of Continua. Then, I created VOT continua in Praat using a script by (Winn, 2020), based on the selective deletion method (Andruski et al., 1994). Each continuum spanned 9 steps from -70 ms to $+90$ ms, incrementing by 20 ms. These VOT boundaries were chosen to fully capture voicing contrasts for bilabial, alveolar, and velar stops across the three target languages. For pitch normalization, F0 was set to 100 Hz across all stimuli to minimize prosodic variation while maintaining natural-sounding male speech.

When automated outputs failed to preserve perceptual naturalness (e.g., inaudible bursts), I manually edited the original voiced productions by extracting the natural burst segment and inserting extended VOT portions from the corresponding voiceless tokens. All concatenations were performed at zero crossings to maintain natural burst intensity and smooth formant transitions. This was especially useful for creating ambiguous and intermediate tokens. The same methodology was applied across all three place-of-articulation continua: /ba/-/pa/, /da/-/ta/, and /ga/-/ka/.

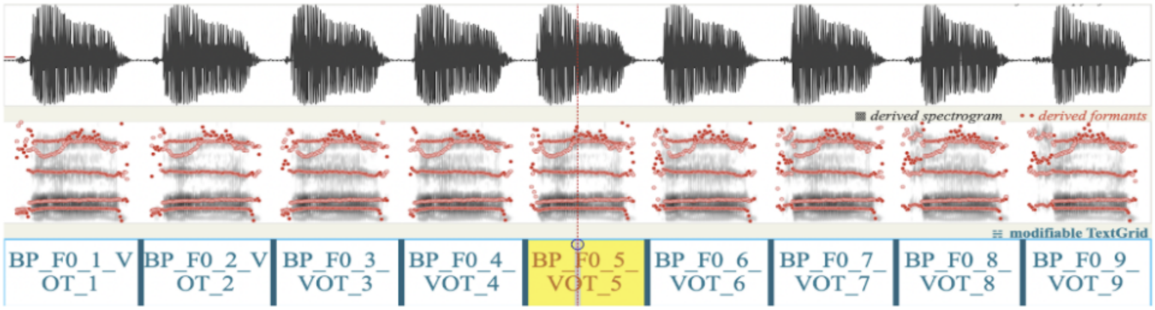


Figure 3.3: Spectrogram and waveform display of the /ba–pa/ continuum (steps 1–9) generated in Praat. Each token increases in voice onset time (VOT) from fully voiced to fully voiceless, illustrating the incremental manipulation at 20 ms per step of the bilabial stop contrast.

Table 3.2: Summary of VOT step number and their corresponding values in milliseconds.

Step	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
VOT (ms)	-70	-50	-30	-10	10	30	50	70	90

Visual Analogue Scale (VAS)

Participants rated each auditory token on a horizontal VAS bar with endpoints labeled by prototypical voiced (on the left) and voiceless categories (right) (e.g., “da” vs. “ta”). The VAS allows participants to respond in a gradient manner, capturing subtle phonetic categorization shifts and certainty levels and rater confidence. Compared to binary forced-choice tasks (e.g., AX or 2IFC), the VAS provides a more sensitive measure of cue weighting and perceptual boundary location. Prior work (Kutlu et al., 2022a) has shown that VAS responses better reflect real-world ambiguity and variability in speech perception, especially when combined with social context. Participants click anywhere on the scale to indicate how they perceive each sound. After each click, the cursor automatically resets to the center to minimize carryover effects or bias from previous responses (e.g., avoiding lazy clicks in the same position across trials).

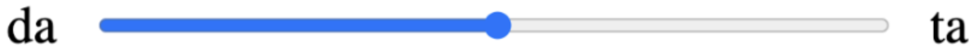


Figure 3.4: VAS interface used in the experiment during a /da/–/ta/ block. The cursor is shown in the center position, where it appears by default at the start of each trial.

3.2.3 Participants

A total of 85 participants were recruited, all aged 18 or older. Inclusion criteria required participants to be dominant in either Russian or Mandarin and to have sufficient English reading proficiency to follow instructions. All participants had lived in an Anglophone country for at least one year and reported no known hearing impairments. Due to limitations in recruiting Russian-dominant speakers on Georgetown’s campus (exacerbated by geopolitical events in recent years), most of our Russian participants were non-students, whom I recruited through community outreach at two Russian Orthodox churches in the D.C. area, word of mouth, and an online form that screened for demographics and language background.

3.2.4 Procedure

Most sessions ($n = 74$) were conducted in the Linguistics Lab on Georgetown University's campus using a laptop connected to an external monitor. For participants unable to commute, I arranged off-campus sessions ($n = 8$). Within the Russian group: two participants were tested at a quiet, not yet opened hair salon above a Russian grocery store, two others were tested at their home. Within the Chinese group, two participants were tested in a quiet church classroom, and two at home of the participants. The same external monitor was used across all sessions to ensure consistent experience and simulation dimensions across all participants.

The experiment began with a short visual practice round and two auditory practice trials (non-target contrasts: /ma/-/la/ and /s/-/ʃ/). Participants then completed the three phases (training, baseline, social) of the experiment. Afterward, we conducted a semi-structured interview lasting $5 - 12$ minutes, depending on participant engagement. I also administered an English oral proficiency task via Elicited Imitation task (EIT) to enable follow-up analyses examining how individuals' L2 English proficiency may correlate with their responses in VAS task. The entire experiment session took around 30 minutes. All procedures were approved by the IRB (#00007321) and designed for accessibility and comfort.

3.3 Statistical Analysis

3.3.1 Data Overview

I collected data from a total of 82 second language users of English: (1) 43 in the Chinese group and 39 in the Russian group. One Russian participant (ID: `028F_RU_AB`) was excluded due to a clear misunderstanding of task instructions. Instead of using the visual analog scale (VAS) to indicate gradient perception, this participant interpreted small movements from the center as categorical choices. As a result, her responses clustered tightly around the center of the scale, in contrast to the more distributed responses from other participants (see Figure B.2 in Appendix B for a comparative visualization). One Chinese participant is excluded due to task incompletion. All other participants' data were retained, yielding a final sample of 42 Mandarin-English bilinguals and 38 Russian-English bilinguals that will be used for downstream analysis.

Table 3.3: Participant breakdown by group, gender, and age.

L1	Gender	Count	Mean Age	Std. Dev.
Chinese listeners	Female	25	26.8	9.0
	Male	17	30.4	13.2
Subtotal		*42*	*28.3*	*10.7*
Russian listeners	Female	26	44.0	14.5
	Male	12	34.5	10.1
Subtotal		*38*	*41.0*	*13.1*
Overall Total		**80**	**34.3**	**11.8**

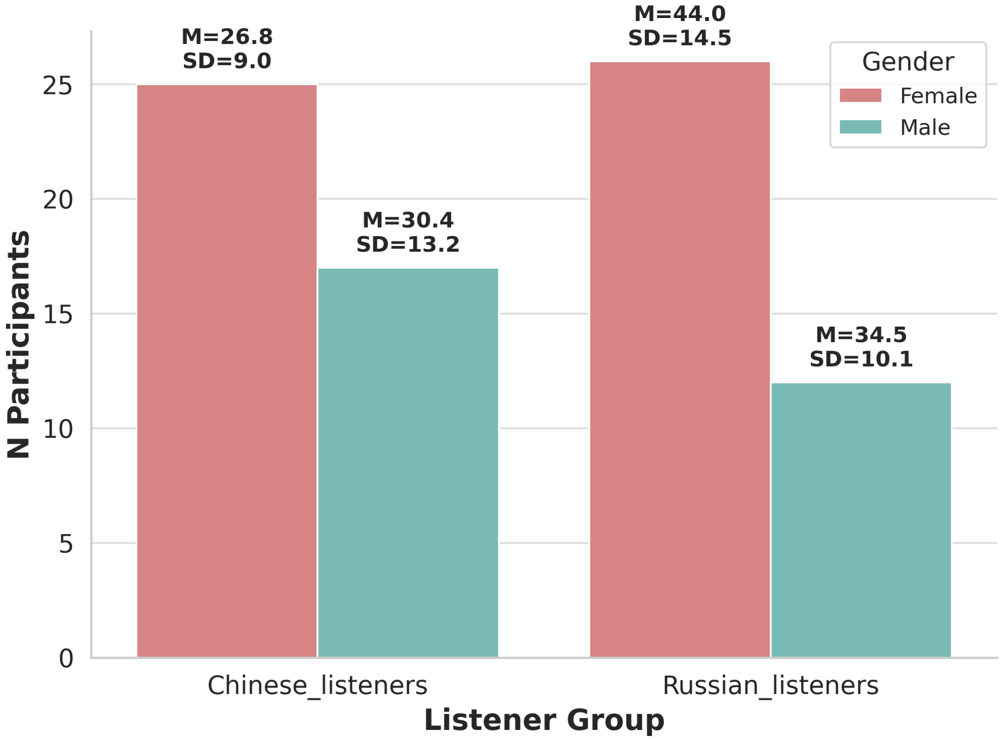


Figure 3.5: Participant demographics (counts) by group and gender (mean age ± sd).

3.3.2 Variables of interest

The dataset comprises the following variables, which serve as predictors and/or grouping factors in subsequent visualizations and statistical modeling:

1. **Listener Group:**

   (a) *Mandarin-English bilinguals* ($n = 43$)

   (b) *Russian-English bilinguals* ($n = 38$)

2. **Condition** (referred to as **"guises"** in modeling): Each participant completed four experimental blocks, with Block 1 always as Baseline. The order of the remaining three guises (Blocks 2-4) was randomized across participants.

(a) *Baseline* (no social information)

(b) Mandarin (Mandarin guise)

(c) *Russian* (Russian guise)

(d) *American* (American English guise)

3. **Block Number**: Indicates the sequence of presentation (1 = *Baseline*, $2-4$ = *guises in random order*).

4. **Step:**

   (a) A categorical variable (1-9) indicating the VOT step along each /ba/-/pa/, /da/-/ta/, or /ga/-/ka/ continuum.

   (b) Step 1 corresponds to a perceptually voiced sound, and Step 9 is voiceless.

5. **POA (Place of Articulation):**

   (a) Bilabial (/ba/-/pa/)

   (b) Alveolar (/da/-/ta/)

   (c) Velar (/ga/-/ka/)

6. **Rating (**referred to as **“Value”** in the models**):** The participant’s click location on the VAS, recorded as a continuous value from 0 to 100. A value of 0 corresponds to the leftmost (voiced) endpoint, and 100 to the rightmost (voiceless) endpoint. Intermediate values reflect gradient judgments.

These variables provide the structure for all downstream visualizations and statistical models, enabling both group-level and condition-level analyses across the VOT continuum and places of articulation.

3.3.3 Analytical Pipeline

This study employed a sequential, model-driven analytical workflow where methodological choices for each research question were empirically informed by prior results. This approach follows best practices in speech perception research (Kleinschmidt and Jaeger, 2015; Wade et al., 2021), ensuring transparency about how analytical decisions emerged from the data. Accordingly, the following section is organized by research question, with each subsection outlining the specific analytical strategies and procedures used to address that question, while maintaining logical connections between stages and their theoretical rationales.

RQ1: Baseline VOT boundaries

Our first research question examines whether Russian-English and Mandarin-English bilinguals maintain L1-typical perceptual boundaries when categorizing voicing in L2 English. To address this, I analyzed baseline block responses using three complementary metrics: (1) mean VAS ratings across the nine VOT steps (−70 ms to 90 ms), (2) the percentage of responses exceeding the 75% voiceless threshold, and (3) the first VOT value where mean ratings crossed this 75% boundary. The 75% threshold was selected to capture intentional voiceless judgments, as values above this threshold require deliberate placement toward the voiceless endpoint of the scale, reducing ambiguity from midline-adjacent clicks that might reflect uncertainty or motor noise.

Patterns in Voiceless Perception. I first observe the percentage of responses reported as voiceless across both listener groups. Between-group differences emerged clearly in the percentage of voiceless responses (See Table 3.4). For alveolar stops at 30 ms, Russian listeners—whose L1 has short-lag VOT—categorized the stimulus as voiceless 41.2% of the time, while Mandarin listeners (with long-lag L1 VOT) did so in only 7.9% of trials. This divergence reflects their L1-specific perceptual tuning: Russian listeners, accustomed to shorter VOT values for voiceless stops, accepted 30 ms as voiceless more readily than Mandarin listeners, who typically require longer VOTs.

Place of articulation further modulated these effects. At 30 ms VOT, bilabial stops were categorized as voiceless most readily, with both groups exceeding 90% voiceless responses. However, for alveolar and velar stops, the same 30 ms VOT was rejected as voiceless by most Mandarin listeners ($< 8\%$ acceptance), while Russian listeners showed moderate acceptance (61.4% for velars). By 50 ms, both groups converged toward categorical voiceless perception ($> 94\%$ acceptance), suggesting that while their boundary locations may differ, their endpoint categorization is similar. These findings also align with aerodynamic evidence that stops with bursts released at a more fronted position (e.g., lips) naturally require shorter VOTs than those produced further back in oral anatomy (e.g., alveolars or velars; (Cho and Ladefoged, 1999)).

Table 3.4: Percentage of voiceless responses (VAS $\geq$ 75) at key VOT points.

L1 Group	**Place of Articulation**	**30 ms**	**50 ms**	**70 ms**	**90 ms**
Mandarin listeners					
	Bilabial	92.1%	98.4%	96.8%	99.2%
	Alveolar	7.9%	94.4%	97.6%	98.4%
	Velar	7.1%	78.6%	98.4%	98.4%
Russian listeners					
	Bilabial	97.4%	98.2%	99.1%	99.1%
	Alveolar	41.2%	94.7%	97.4%	99.1%
	Velar	61.4%	97.5%	98.2%	100.0%

Note. The values represent the percentage of responses $\geq$ 75 on the VAS scale $(0 - 100)$, where 100 means completely voiceless.

Boundary Shifts Relative to L1 Norms. I define each group's perceptual boundary as the first VOT step (reported as durational value in ms) at which their mean VAS rating reaches or exceeds 75% "voiceless." Whereas Table 3.4 showed the percentage of individual responses crossing that threshold, this boundary metric abstracts away trial-to-trial noise to pinpoint where the group's category judgment becomes reliably voiceless. Table 3.5 juxtaposes two sets of values:

1. **Published monolingual production norms.** Left columns reflect the *average VOTs produced* by native speakers of Russian (short-lag), English (mid-lag), and Mandarin (long-lag) for voiceless stops in each place of articulation.

2. **Observed bilingual perceptual boundaries.** Right columns show the 75 % crossover points from our baseline block for Mandarin-English and Russian-English bilingual listeners.

Table 3.5: VOT boundaries for voiceless perception: published norms versus observed.

Place of Articulation	Reported VOT (ms)			Observed (VAS $\geq$ 75)	
	Russian	English	Mandarin	Mandarin	Russian
Bilabial	18	58	82	30	30
Alveolar	20	70	81	50	50
Velar	38	80	92	50	50

Note. Published VOT norms are from (Ringen and Kulikov, 2012), (Lisker and Abramson, 1964), and (Chao and Chen, 2008). Observed boundaries correspond to the 75% voicelessness threshold derived from VAS ratings in the baseline condition.

Despite earlier group differences in raw voiceless response rates (See Table 3.4), both bilingual groups set identical categorical boundaries, 30 ms for bilabials and 50 ms for alveolars and velars. This convergence is expected, as once mean ratings pass the 75% mark, individual variability has been largely smoothed out by averaging.

In Figure 3.6, I overlay these observed boundaries (solid green lines) on each group's mean rating curves (± SE), with Mandarin-English listeners in the top row and Russian-English listeners in the bottom row. The dashed vertical lines mark the published production norms for Russian (blue), English (orange), and Mandarin (red).

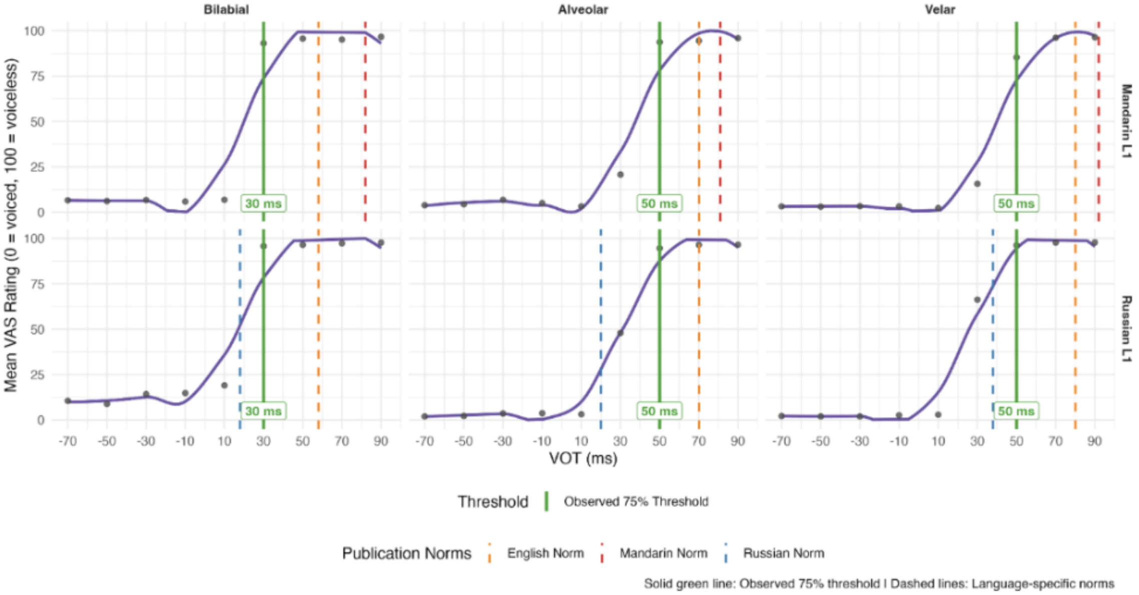


Figure 3.6: Bilinguals' voiceless perception identification at 75% threshold. Comparison of bilingual responses to L1 and English phonetic norms.

For the Mandarin–English group (top row), the green line falls well to the left of both the Mandarin (long-lag) and English (mid-lag) dashed lines in every panel, indicating that these listeners require less positive VOT than either the reported native group to reach a reliable "voiceless" judgment. In contrast, the Russian–English group's green line (bottom row) consistently falls between the Russian (short-lag) and English (mid-lag) dashed lines, showing that Russian bilinguals' default voiceless perception is already shifted upward from their L1's shorter VOT toward English.

Across all three places of articulation, both bilingual groups gravitate toward English-like VOT categories. Mandarin bilinguals overshoot in the opposite direction (undershooting both L1 and L2 norms), whereas Russian bilinguals converge into the English range but do not fully reach the English mean.

RQ 2: Does social information about the speaker (social cue) modulate shifts from baseline?

Before asking which guises shift voicing perception (RQ 3), I must first show that any social information departs from baseline (RQ 2) systematically. Because each listener hears three different guises in succession, later blocks may reflect both carry-over from earlier exposure and uneven guise assignment. To isolate the initial impact of talker identity, I therefore compare only the Baseline block (no guise) to each listener's First Social block (first introduction of a guise). The sections follow first offer three pieces of evidence, from methodological and theoretical considerations, as to why comparing baseline vs. first social block responses is the best approach to the current questions; then, I show statistical reports on how the four conditions—baseline and the three social guises—interact with VOT step in shaping bilingual listeners' voicing contrast perception.

Block Order Effect. I begin by collapsing across places of articulation to plot the full distribution of slider ratings in each of the four blocks (`Baseline` → `First_Social` → `Second_Social` → `Third_Social`), faceted by the listener `group` (Figure 3.7). Mandarin–English bilinguals show only minor shifts from `Baseline` into each `Social block`, indicating relative stability in their categorization over time. Russian–English bilinguals exhibit a modest rise at `First_Social` that grows in `Second_Social` before partially receding in `Third_Social`.

These patterns collapse over PoA and do not account for the fact that guises were assigned unevenly across blocks. Their relative stability nonetheless hints that the largest social-cue effect averaged across both groups occurs at `First_Social` (the first divergence from `Baseline`), motivating our focus on that contrast, and also suggests that between listener groups, Russian-English bilinguals may be more prone to social effects than the Mandarin-English group.

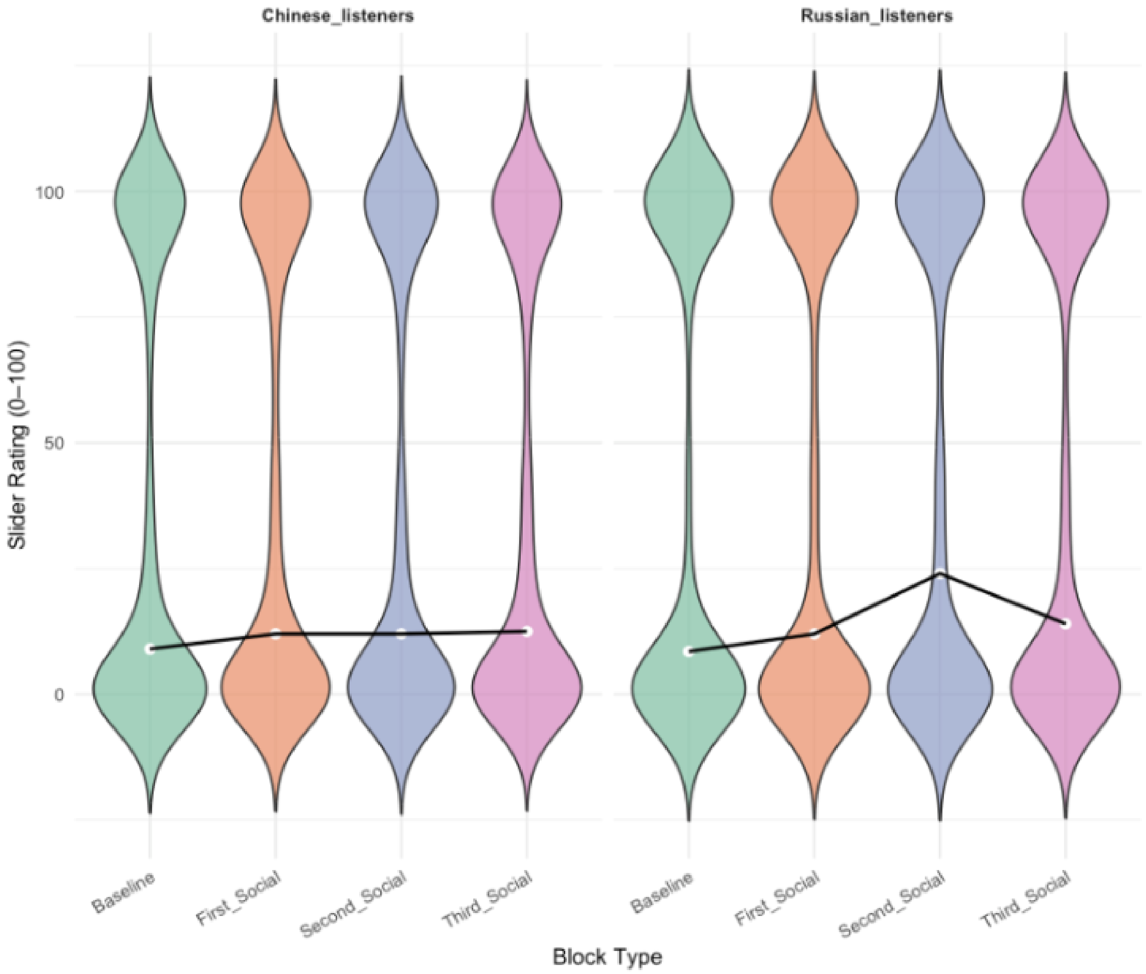


Figure 3.7: Distribution of slider ratings by block order across both listener groups.

Block-Order* × *Guise Imbalance (Chi-square Tests). To confirm that guise assignment varied systematically by block, I constructed a 3 × 3 count table of `Block_Type` (`First_Social / Second_Social / Third_Social`) by `Guise_Type`

(American / Mandarin / Russian) for each group and ran Pearson's chi-square tests of independence. See Table 3.6 for count distributions across guises and blocks.

Table 3.6: Counts of social guise by block type for Mandarin and Russian bilinguals.

Listener Group	Block Type	American	Mandarin	Russian
Chinese listeners	First_Social	1053	972	1377
	Second_Social	1377	897	972
	Third_Social	972	1377	1053
Russian listeners	First_Social	1134	1053	891
	Second_Social	891	966	1053
	Third_Social	1053	891	1145

Note. Each value represents the count of trials per guise and block type for the corresponding listener group.

Chinese listeners: $\chi^2(4) = 281.02$, $p < .001$. Standardized residuals showed pronounced imbalances: in First_Social, Russian guise was over-represented ($z = +10.04$) while American ($z = -4.39$) and Mandarin ($z = -5.72$) were under-represented; in Second_Social, American was over-represented ($z = +12.54$) and Mandarin ($z = -6.91$) and Russian ($z = -5.72$) were under-represented; in Third_Social, Mandarin was over-represented ($z = +12.54$) and American ($z = -8.00$) and Russian ($z = -4.39$) under-represented.

Russian listeners: $\chi^2(4) = 66.60$, $p < .001$. Here, First_Social showed over-representation of the American ($z = +4.17$) and Mandarin ($z = +3.09$) guises, alongside under-representation of the Russian guise ($z = -7.21$); Second_Social under-represented American ($z = -4.61$) while Russian was over-represented ($z = +3.09$); Third_Social over-represented the Russian guise ($z = +4.17$) and under-represented Mandarin ($z = -4.61$).

These significant imbalances (See Figure 3.8) confirm that `Block_Type` and `Guise_Type` were not independent for either group, motivating our focus on the `Baseline` versus each participant's `First_Social` block for a clean between-subjects comparison in RQ 2.

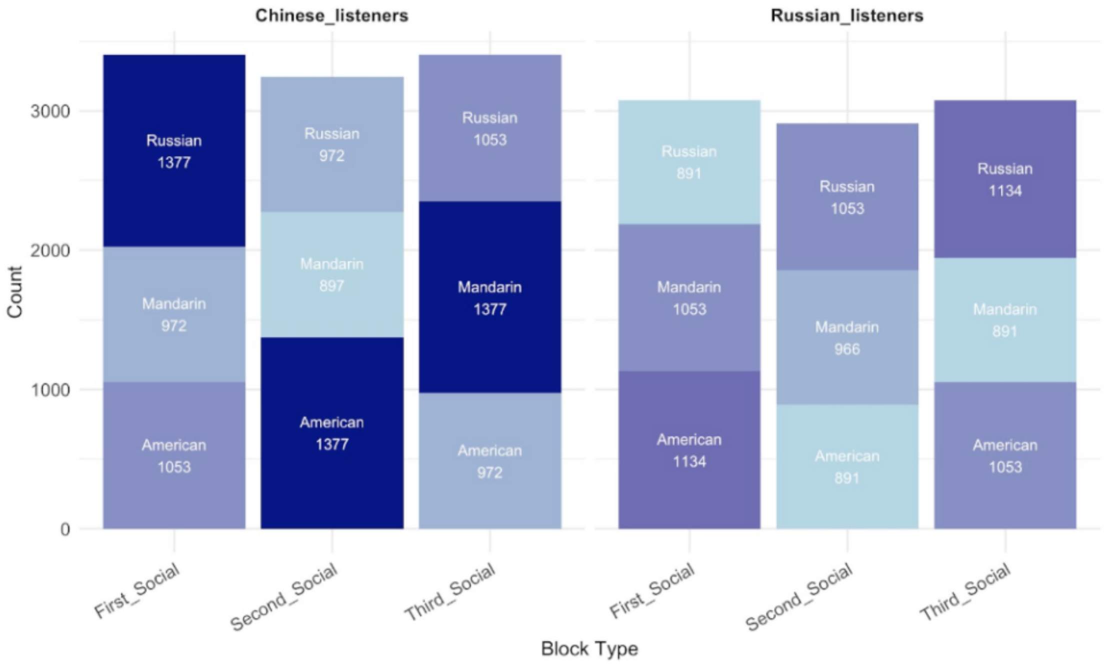


Figure 3.8: Gradient-filled stacked barplot on guise distribution by block order. Guise counts for trials by block type and listener group are shown. The numbers in each box refer to the total number of trials available for that block.

The order effects and imbalance between guise assignment (as randomly generated by PCIbex) further support the analytical approach to restrict RQ 2 analyses to **Baseline** vs. **First_Social** only.

Theoretical Justification: Accumulation of Social Knowledge. Statistical imbalance aside, social perceptual learning is a dynamic process, whereby initial exposure often differs qualitatively from later experience. (Zellou et al., 2017) used a shadowing paradigm to show that listeners cumulatively build an internal model of a talker's articulatory style. Participants gradually shifted their own nasality convergence across experimental blocks as they accumulated exposure, even when the talker's speech characteristics varied unpredictably. Similarly, (Goldinger, 1998) demonstrated that early-phase lexical decision tasks are strongly influenced by the first few exposures to a voice, with diminishing returns over time.

In our perception task, each social block adds another layer of "talker" expectation, which listeners carry forward their evolving beliefs about how a given guise should sound. Comparing all three social blocks at once would wash out the initial effect of introducing a new social cue, obscuring how any given social exposure shapes voicing judgments. By focusing on **Baseline** versus **First_Social**, we capture the cleanest window into talker-identity effects on speech perception, free from the confounds of cumulative social learning. Having justified our focus on the Baseline vs. First Social contrast, I now ask whether talker guises exert a statistically reliable effect on English VOT perception for either listener group.

Global Interactions. The preceding section our decision to compare the baseline block with each listener's first social block, arguing that this contrast offers (i) the cleanest temporal window on the initial impact of talker identity, (ii) freedom from the guise-by-block imbalances documented in the chi-square analyses, and (iii) theoretical alignment with work showing that listeners' perceptual expectations update rapidly after the very first exposure to a talker (Zellou et al., 2017). What follows reports the statistical tests that address RQ 2 proper: **Do guises exert any measurable influence on VOT perception once the experiment introduces social information?** I proceed from the most general result to the most localised, finishing with a short summary that sets up RQ 3.

Global interaction models. I first asked whether the presence of talker identity matters at all, irrespective of PoA. For each listener group, a linear mixed-effects model fitted with Condition (Baseline, American, Mandarin, Russian), VOT step (nine-level factor, or step_f in the model), and their interaction as fixed factors. I included participants IDs and place of articulation (PoA) as random intercepts to capture between-speaker and articulatory variability. Table 3.7 reports the omnibus Type III Wald χ^2 tests for the fixed effects. In both bilingual groups the Condition × step interaction is highly significant (Mandarin L1: $\chi^2 = 67.9$; Russian L1: $\chi^2 = 124.5$, *p*s $< .001$), whereas the Condition main effect is not. In other words, talker identity never causes a blanket up- or down-shift in slider ratings; its influence is confined to particular VOT steps.

Table 3.7: Type III Wald χ^2 tests for global interaction models.

Listener Group	**Effect**	$\boldsymbol{\chi^2}$	**df**	**p-value**
Mandarin–English	Condition	2.3398	3	.505
	step_f	20,619.9670	8	$< .001^{***}$
	Condition × step_f	67.8757	24	$< .001^{***}$
Russian–English	Condition	1.4545	3	.693
	step_f	15,680.0685	8	$< .001^{***}$
	Condition × step_f	124.5225	24	$< .001^{***}$

Note. Type III Wald χ^2 tests correspond to fixed effects in the mixed-effects model. Asterisks denote significance levels ($^{***}p < .001$).

Both groups show the expected, very large main effect of step_f (Chinese: $\chi^2 = 20.620$, *df* $= 8$; Russian: $\chi^2 = 15.680$, *df* $= 8$; both $p < .001$). More critically, the **Condition** × step_f interaction reaches significance in both groups (Chinese: $\chi^2 = 67.88$, *df* $= 24$, $p < .001$; Russian: $\chi^2 = 124.52$, *df* $= 24$, $p < .001$), whereas the Condition main effect alone is non-significant (*p*s $> .50$). The absence of an overall Condition main effect tells us there is no uniform up- or down-shift in ratings across the entire continuum. Instead, the significant interaction indicates that any social influence is confined to particular VOT steps. To learn which ones, and whether the effect generalizes across PoAs, I turn to PoA-specific models.

What is shifting? Mandarin–English bilinguals show a single, focused effect: the Russian guise boosts voiceless responses at Step 6 by approximately ≈ +8 points (Table 3.8).

Russian–English bilinguals show a richer, bidirectional pattern: the Mandarin guise nudges ratings upward at Step 5 but reduces them from Step 7 onward, while the Russian guise again raises ratings at Step 6 (Table 3.9).

Table 3.8: Significant fixed effects in global model: Mandarin–English bilinguals.

Term	Estimate	Std. Err.	t value	p value
`step_f6`	38.60	1.21	31.89	$< .001^{***}$
`step_f7`	87.01	1.21	71.89	$< .001^{***}$
`step_f8`	90.67	1.21	74.92	$< .001^{***}$
`step_f9`	91.73	1.21	75.79	$< .001^{***}$
`ConditionRussian : step_f6`	8.45	2.25	3.75	$< .001^{***}$

Note. Significant fixed effects from the global linear mixed-effects model for Mandarin–English bilinguals. Asterisks denote significance levels ($^{***}p < .001$).

Significant Effects in Global Model for Chinese. The interaction model confirms a highly categorical VOT response: compared to Step 1, slider ratings at Steps 6-9 are massively higher ($\beta s = 38.60 - 91.73$, $t > 31.9$, $p < .001$), reflecting the sharp perceptual boundary toward voiceless. Crucially, social-cue effects emerge only at that boundary: the Russian guise produces an additional +8.45-point increase in voiceless ratings specifically at Step 6 ($t = 3.75$, $p < .001$). No other `Condition` × `step` contrasts reach significance, indicating that talker identity exerts its influence precisely at the most ambiguous VOT region.

Table 3.9: Significant fixed effects in the global model – Russian–English.

Term	Estimate	Std. Err.	t value	p value
`step_f5`	3.48	1.44	2.42	.016
`step_f6`	65.08	1.44	45.32	$< .001^{***}$
`step_f7`	90.89	1.44	63.29	$< .001^{***}$
`step_f8`	92.28	1.44	64.26	$< .001^{***}$
`step_f9`	92.49	1.44	64.41	$< .001^{***}$
`ConditionMandarin : step_f5`	6.14	2.84	2.16	.031
`ConditionRussian : step_f6`	8.43	3.03	2.78	.005
`ConditionMandarin : step_f7`	-7.36	2.84	-2.59	.010
`ConditionMandarin : step_f8`	-10.49	2.84	-3.69	$< .001^{***}$
`ConditionMandarin : step_f9`	-10.46	2.84	-3.68	$< .001^{***}$

Note. Significant fixed effects from the global linear mixed-effects model for Russian–English bilinguals. Asterisks denote significance levels ($^{***}p < .001$).

Significant Effects in Global Model for Russians. As with Chinese bilinguals, Russian listeners display strong categorical effects across the VOT continuum: Steps 5-9 all differ reliably from Step 1 ($\beta s = 3.48 - 92.49$, $ps \leq .016$). Crucially, social-cue × step interactions emerge at multiple ambiguity points: under the Mandarin guise, ratings shift by +6.14 points at Step 5 ($t = 2.16$, $p = .031$), -7.36 at Step 7 ($t = -2.59$, $p = .010$), and $-10.49/-10.46$ at Steps 8-9 (ts $\approx -3.69/-3.68$, ps $< .001$). Under the Russian guise, ratings increase by +8.43 points at Step 6 ($t = 2.78$, $p = .005$). These effects indicate that Russian bilinguals' boundary shifts vary in directionality depending on VOT and perceived talker identity: they tune differently to Mandarin versus Russian guises at specific perceptual boundary regions.

Place-of-articulation models. Because English VOT lags vary by PoA, the interaction model is reassessed separately for bilabial (BP), alveolar (DT), and velar (GK) stops. Only `Condition` × `step` terms with $p < .05$ are summarized in Tables 3.10- 3.11. From each model, only those fixed-effect terms whose p-value $< .05$ are extracted; for brevity, I omit step main effects, which confirm categorical rise in voiceless judgments.

For Mandarin–English bilinguals, social identity effects on voicing perception are extremely focal: the Russian guise yields a +8.45-point increase in slider ratings only at the ambiguous VOT step (Step 6), and this holds for bilabial, alveolar, and velar stops alike. No other `guise` × `step` combinations reached significance, indicating that talker identity only modulates perception at the precise boundary region.

Table 3.10: Significant `condition` × `step_f` interactions in PoA-specific models – Mandarin–English bilinguals.

PoA	Interaction	Estimate	Std. Err.	t value	p value
BP	`ConditionRussian : step_f6`	8.45	2.35	3.59	$< .001^{***}$
DT	`ConditionRussian : step_f6`	8.45	2.35	3.59	$< .001^{***}$
GK	`ConditionRussian : step_f6`	8.45	2.35	3.59	$< .001^{***}$

Note. Each model was estimated separately by place of articulation (PoA). Asterisks denote significance levels ($^{***}p < .001$).

Russian listeners display a broader pattern of social-cue modulation at the perceptual boundary region. At Step 5 (just before the boundary shift), the Mandarin guise raises voiceless ratings by $\approx +6$ points ($p \approx .036$) for every PoA, whereas the Russian guise is not yet reliable. At Step 6 the pattern shifts: the Russian guise now contributes an $\approx +8$ point boost ($p \approx .007$), while the Mandarin contrast is null. Beyond the boundary (Steps 7-9), the Mandarin guise reverses direction, lowering ratings by $\approx 7 - 10$ points (all ps $\leq .012$). These bidirectional effects recur across bilabial, alveolar, and velar stops, indicating that Russian listeners' perceptual boundary is flexibly reshaped by talker identity in both directions, depending on the social cue.

Table 3.11: Significant condition × step_f interactions in PoA-specific models – Russian–English bilinguals.

PoA	Interaction	Estimate	Std. Err.	t value	p value
BP	ConditionMandarin : step_f5	6.14	2.93	2.09	.036
BP	ConditionRussian : step_f6	8.43	3.12	2.70	.007
BP	ConditionMandarin : step_f7	-7.36	2.93	-2.51	.012
BP	ConditionMandarin : step_f8	-10.49	2.93	-3.58	$< .001^{***}$
BP	ConditionMandarin : step_f9	-10.46	2.93	-3.57	$< .001^{***}$
DT	ConditionMandarin : step_f5	6.14	2.93	2.09	.036
DT	ConditionRussian : step_f6	8.43	3.12	2.70	.007
DT	ConditionMandarin : step_f7	-7.36	2.93	-2.51	.012
DT	ConditionMandarin : step_f8	-10.49	2.93	-3.58	$< .001^{***}$
DT	ConditionMandarin : step_f9	-10.46	2.93	-3.57	$< .001^{***}$
GK	ConditionMandarin : step_f5	6.14	2.93	2.09	.036
GK	ConditionRussian : step_f6	8.43	3.12	2.70	.007
GK	ConditionMandarin : step_f7	-7.36	2.93	-2.51	.012
GK	ConditionMandarin : step_f8	-10.49	2.93	-3.58	$< .001^{***}$
GK	ConditionMandarin : step_f9	-10.46	2.93	-3.57	$< .001^{***}$

Note. Each model was estimated separately by place of articulation (PoA). Asterisks denote significance levels ($^{***}p < .001$).

These PoA-specific model outputs show that while both groups are sensitive to talker identity at the categorical boundary, Russian listeners exhibit a more complex pattern, likely reflecting their richer L1 phonetic space and greater flexibility in weighting social cues across the VOT continuum.

To complement the tables, mean slider ratings are plotted for all four conditions across the nine VOT steps, faceted by PoA (columns) and listener group (rows). Figure 3.9 overlays mean slider ratings for the four conditions, faceted by PoA. In the `Mandarin` group, we can see that the guises didn't produce any clear directions for most steps, except at Step 6 (the most ambiguous step), where `Russian` prompted the most voiceless rating, followed `American` and `baseline`, then lastly `Mandarin`.

This effect was more obvious for the alveolar pair than other PoAs. Russian listeners show a more intricate shape: an early Mandarin rise (Step 5), a Russian-guise bump at Step 6, followed by a Mandarin dip through Steps 7-9—again consistently across the three PoAs. The pattern holds true across all places of articulation. Moreover, the Russian group also showed consistent guise differences in earlier VOT steps (e.g., Steps 1-5; although not statistically significant), whereby a Chinese guise suppressed voiced categorization and Russian guise strengthened voiced perception, especially in bilabial and alveolar context.

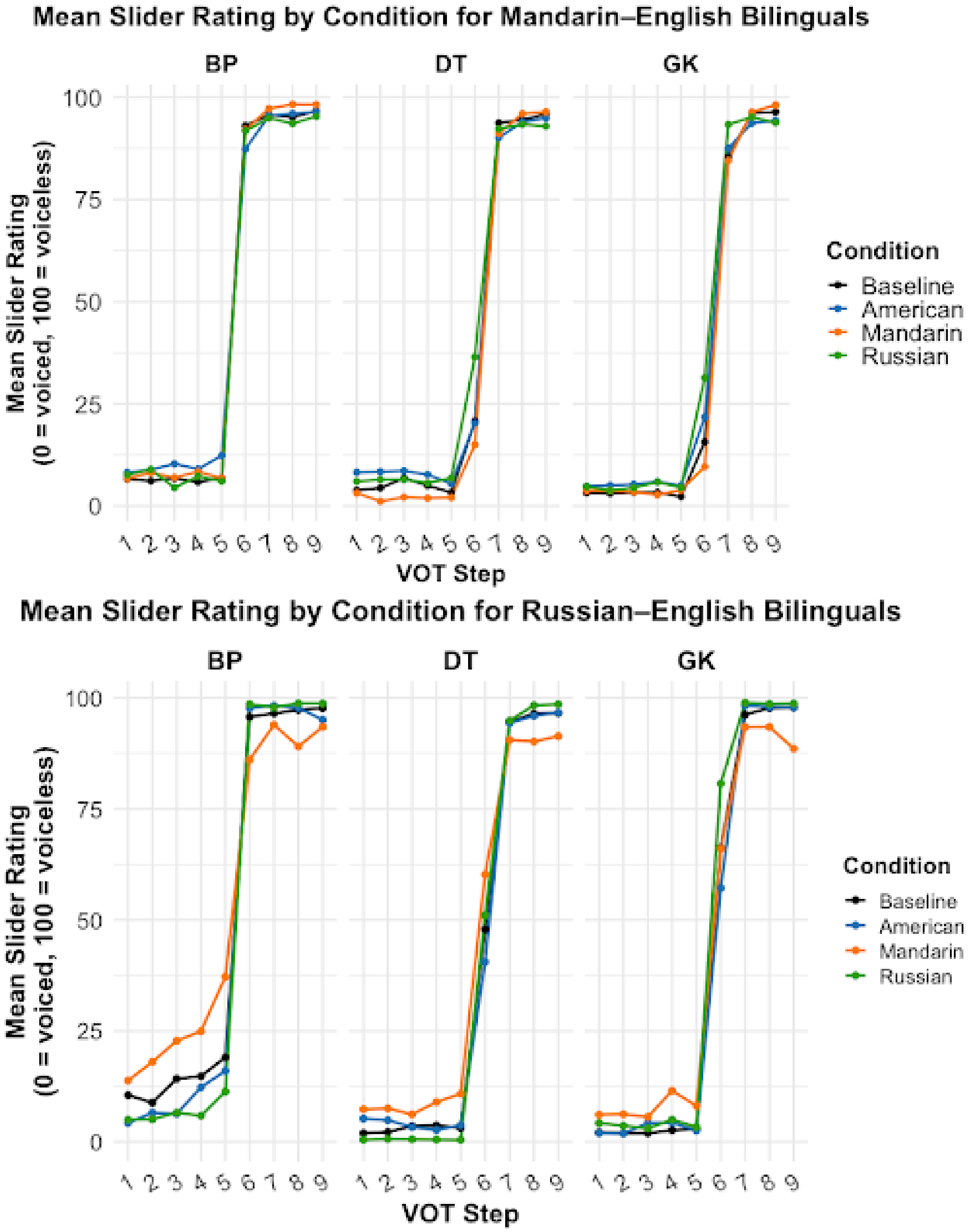


Figure 3.9: Mean slider ratings by listener group and PoA. Top for Mandarin–English bilinguals ($N = 42$), bottom for Russian–English bilinguals ($N = 38$). Baseline = black, American = blue, Mandarin = orange, Russian = green. Top: Mandarin; bottom: Russian.

These observations confirm that social information matters, but only where listeners are uncertain and only to the extent allowed by their own L1-specific perceptual routines. The focal, step-dependent influence documented here provides a clear springboard for the next section, where I examine why the two bilingual groups respond differently and how those differences illuminate the mechanisms of cross-linguistic boundary tuning.

RQ 3: Directionality of Socially-Modulated Perceptual Shifts

Zone Definitions and Directional Framework. To evaluate whether social-induced shifts aligned with phonological predictions, I segmented the VOT continuum into voiced (short VOT) and voiceless (long VOT) zones customized per listener group and place of articulation (Table 3.12). This approach accounts for two critical factors: first, the cross-linguistic differences in voicing boundaries, where Russian exhibits earlier voiceless boundaries (shorter VOT) than English, while Mandarin exhibits later boundaries (longer VOT) than English; and second, the known PoA-specific effects on VOT perception. Our zone definitions (Table 3.12) were empirically derived from categorical boundary shifts observed during data exploration (See Figure 3.9).

Table 3.12: Definition of voiced and voiceless zones by listener and PoA.

Listener Group	PoA	Voiced Zone Steps	Voiceless Zone Steps
Mandarin–English	BP	1-5	6-9
	DT	1-6	7-9
	GK	1-6	7-9
Russian–English	BP	1-5	6-9
	DT	1-5	7-9
	GK	1-5	6-9

Omnibus Social Cue Effects Across Perceptual Zones (Table 3.13). The Type III omnibus χ^2 tests reveal compelling evidence that social guises systematically modulate voicing perception across critical boundary regions. Table 3.13 summarizes significant Condition effects (χ^2, p-values) for all 12 listener group × PoA × zone combinations, with 9 of 12 zones showing statistically significant social-cue effects ($p < .05$).

Table 3.13: Zone-wise omnibus tests of condition.

Listener Group	PoA	Zone	χ^2	df	p-value
Russian	BP	Voiced	1.32	3	0.725
Mandarin	BP	Voiced	11.47	3	0.009
Russian	DT	Voiced	25.23	3	$< .001$
Mandarin	DT	Voiced	12.85	3	0.005
Russian	GK	Voiced	36.28	3	$< .001$
Mandarin	GK	Voiced	18.75	3	$< .001$
Russian	BP	Voiceless	26.95	3	$< .001$
Mandarin	BP	Voiceless	9.66	3	0.022
Russian	DT	Voiceless	12.54	3	0.006
Mandarin	DT	Voiceless	6.85	3	0.077
Russian	GK	Voiceless	9.84	3	0.020
Mandarin	GK	Voiceless	0.29	3	0.961

Directionality of Social Cue Effects: Mandarin–English Bilinguals. Linear mixed-effects models revealed significant social-cue effects across voiced and voiceless zones for Mandarin–English bilinguals, with patterns partially supporting phonological predictions.

In voiced zones (Table 3.14; top), Mandarin–English bilinguals exhibited significant guise effects across places of articulation. For alveolar stops, the Russian guise increased ratings by 13.52 points at Step 6 ($t = 3.96$, $p < .001$). Velar stops showed contrasting effects: the Mandarin guise decreased ratings by 6.59 points at Step 6

(t = -2.03, p = 0.043), while the Russian guise increased them by 14.14 points ($t = 4.96$, $p < .001$). Figure 3.10 visually captures this opposition, with Russian-guise enhancement and Mandarin-guise suppression at the velar boundary.

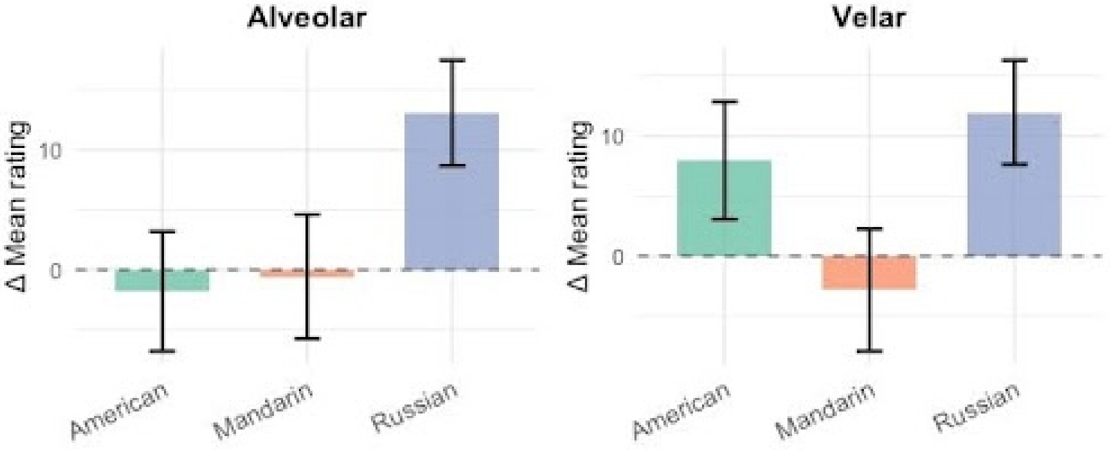


Figure 3.10: Mandarin–English bilinguals – velar voiced zone (Step 6). Delta ratings relative to baseline showing predicted Mandarin-guise suppression and novel Russian-guise enhancement.

Table 3.14: Significant condition and interaction effects – Mandarin–English bilinguals.

Listener	PoA	Zone	Term	Estimate	Std. Err.	t	p
			Voiced Zone				
Mandarin	DT	Voiced	`ConditionRussian : step_f6`	13.52	3.41	3.96	$< .001$
Mandarin	GK	Voiced	`ConditionMandarin : step_f6`	-6.59	3.25	-2.03	0.043
Mandarin	GK	Voiced	`ConditionRussian : step_f6`	14.14	2.85	4.96	$< .001$
			Voiceless Zone				
Mandarin	BP	Voiceless	`ConditionAmerican`	-6.01	1.70	-3.54	$< .001$
Mandarin	BP	Voiceless	`ConditionAmerican : step_f7`	5.59	2.29	2.44	0.015
Mandarin	BP	Voiceless	`ConditionAmerican : step_f8`	6.64	2.29	2.90	0.004
Mandarin	BP	Voiceless	`ConditionAmerican : step_f9`	5.54	2.29	2.42	0.016
Mandarin	DT	Voiceless	`ConditionAmerican`	-3.00	1.50	-2.00	0.046
Mandarin	DT	Voiceless	`ConditionMandarin`	-3.10	1.55	-1.99	0.047
Mandarin	DT	Voiceless	`ConditionMandarin : step_f8`	4.08	2.04	2.00	0.046
Mandarin	GK	Voiceless	`ConditionRussian`	7.09	2.20	3.23	0.001
Mandarin	GK	Voiceless	`ConditionRussian : step_f8`	-9.03	2.97	-3.04	0.002
Mandarin	GK	Voiceless	`ConditionRussian : step_f9`	-10.62	2.97	-3.57	$< .001$

In voiceless zones (Table 3.14; bottom), complex patterns emerged. Bilabial stops showed an overall American-guise suppression ($\beta = -6.01$, $p < .001$), but significant increases at Steps 7-9 (e.g., Step 8: $\beta = 6.64$, $p = 0.004$), visualized in Figure 3.11. Alveolar stops displayed American-guise suppression ($\beta = -3.00$, $p = 0.046$) and Mandarin-guise suppression ($\beta = -3.10$, $p = 0.047$), though the latter increased ratings at Step 8 ($\beta = 4.08$, $p = 0.046$), creating a bifurcated pattern shown in Figure 3.12. Velar stops revealed Russian-guise enhancement overall ($\beta = 7.09$, $p = 0.001$) but suppression at Steps 8-9 (e.g., Step 9: $\beta = -10.62$, $p < .001$), see Figure 3.13.

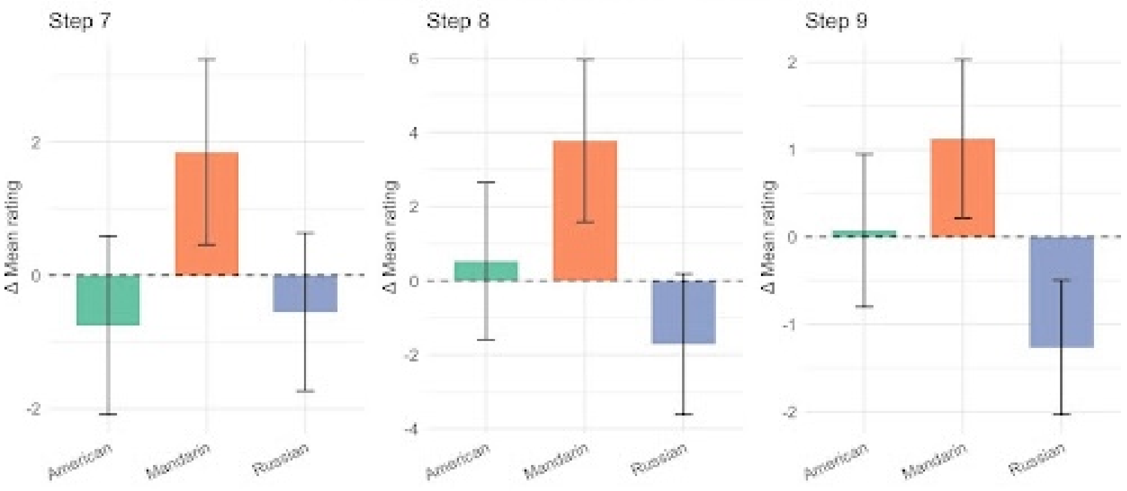


Figure 3.11: Mandarin–English bilinguals – bilabial voiceless zone (steps 7–9). Contradictory American-guise effects reduce voiceless perception, and Mandarin guise does the reverse.

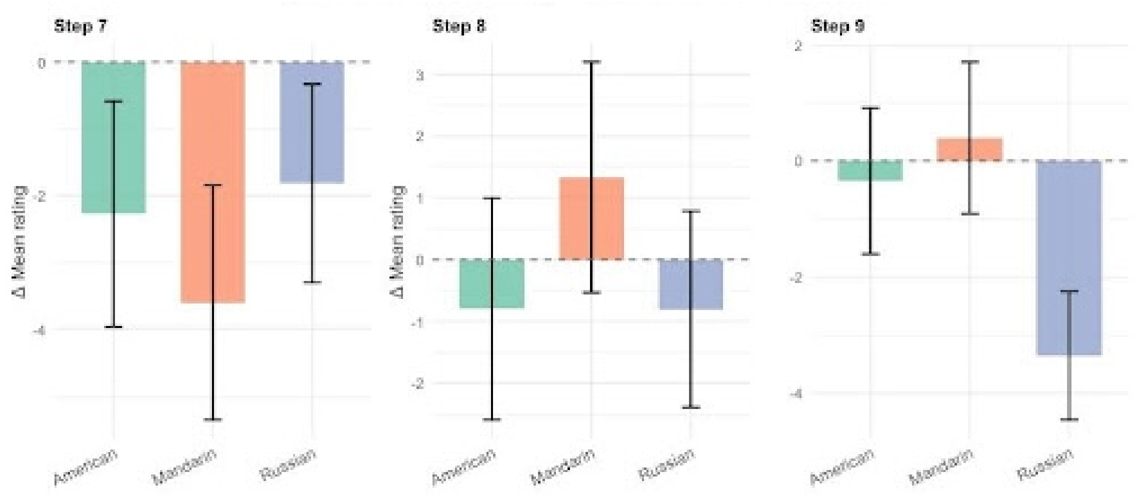


Figure 3.12: Mandarin–English bilinguals alveolar voiceless zone (steps 7–9).

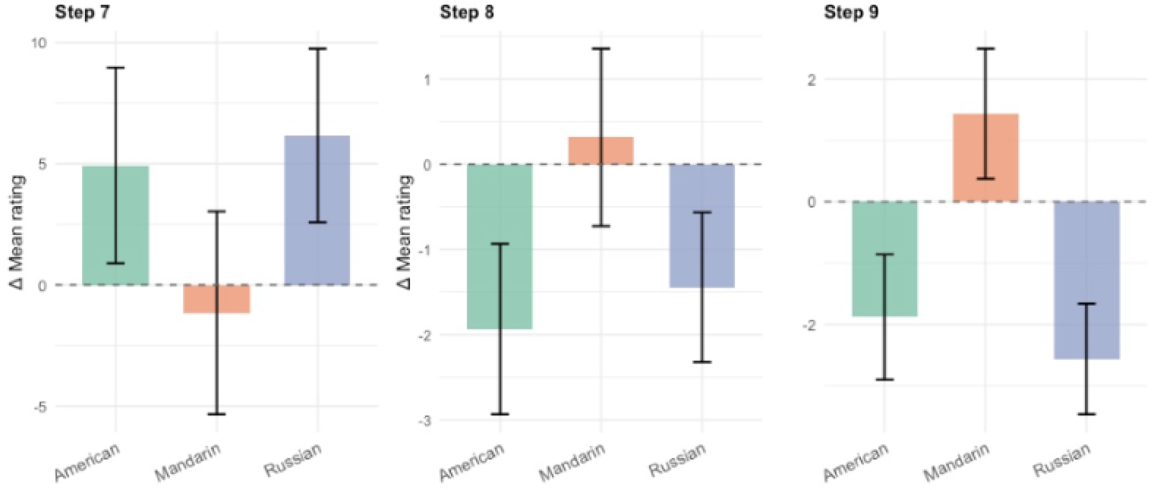


Figure 3.13: Mandarin–English bilinguals velar voiceless zone (steps 7–9).

Directionality of Social Cue Effects: Russian–English Bilinguals. Voiced zones (Table 3.15; top) featured prominent uncertainty amplification. Bilabial stops showed Mandarin guise increases at Step 5 ($\beta = 14.82$, $p = 0.036$), expanding response ranges in Figure 3.14. Alveolar stops exhibited rating increases under both American ($\beta = 3.81$, $p = 0.026$) and Mandarin guises ($\beta = 4.07$, $p = 0.021$), visualized in Figure 3.15. Velar stops displayed Mandarin-guise elevation at Step 4 ($\beta = 4.95$, $p = 0.029$), broadening perceptual ambiguity in Figure 3.16.

Table 3.15: Significant condition and interaction effects – Russian–English bilinguals.

Listener	PoA	Zone	Term	Estimate	Std. Err.	t	p
			Voiced Zone				
Russian	BP	Voiced	`ConditionMandarin : step_f5`	14.82	5.54	2.68	0.008
Russian	DT	Voiced	`ConditionAmerican`	3.81	1.71	2.23	0.026
Russian	DT	Voiced	`ConditionMandarin`	4.07	1.76	2.31	0.021
Russian	GK	Voiced	`ConditionMandarin`	3.58	1.67	2.14	0.032
Russian	GK	Voiced	`ConditionMandarin : step_f4`	4.95	2.27	2.19	0.029
			Voiceless Zone				
Russian	BP	Voiceless	`ConditionMandarin`	-8.76	1.90	-4.61	$< .001$
Russian	BP	Voiceless	`ConditionMandarin : step_f7`	7.10	2.59	2.74	0.006
Russian	BP	Voiceless	`ConditionMandarin : step_f9`	5.39	2.59	2.08	0.038
Russian	GK	Voiceless	`ConditionAmerican`	-7.74	3.90	-1.98	0.047
Russian	GK	Voiceless	`ConditionRussian`	14.43	4.30	3.36	$< .001$
Russian	GK	Voiceless	`ConditionAmerican : step_f7`	11.10	5.35	2.08	0.038
Russian	GK	Voiceless	`ConditionRussian : step_f7`	-11.75	5.86	-2.01	0.045
Russian	GK	Voiceless	`ConditionRussian : step_f8`	-13.50	5.86	-2.31	0.021
Russian	GK	Voiceless	`ConditionRussian : step_f9`	-13.52	5.86	-2.31	0.021

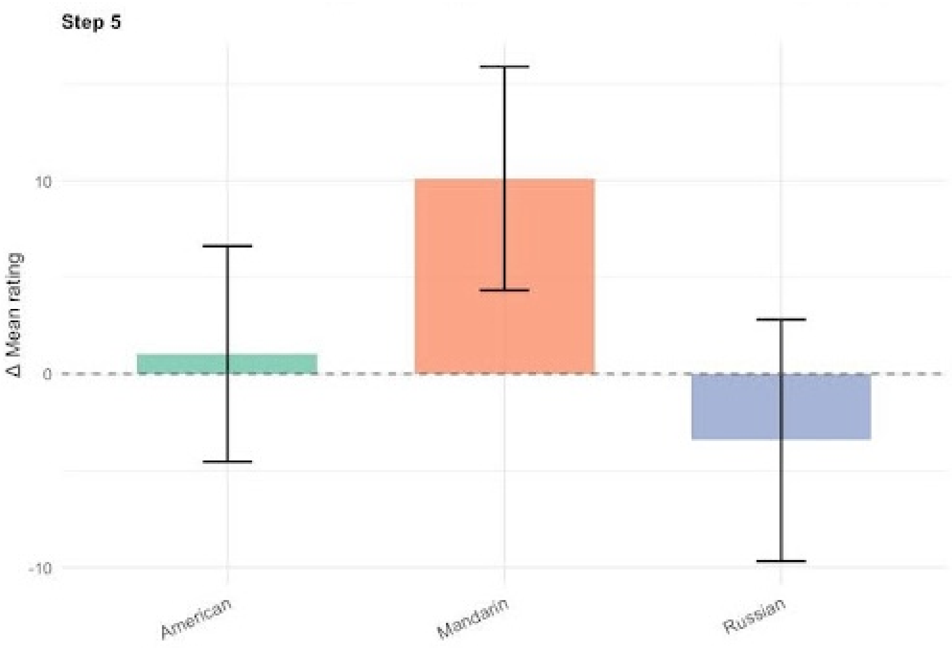


Figure 3.14: Russian–English bilinguals – bilabial voiced zone (step 5). Mandarin guise amplifies perceptual uncertainty at the bilabial boundary.

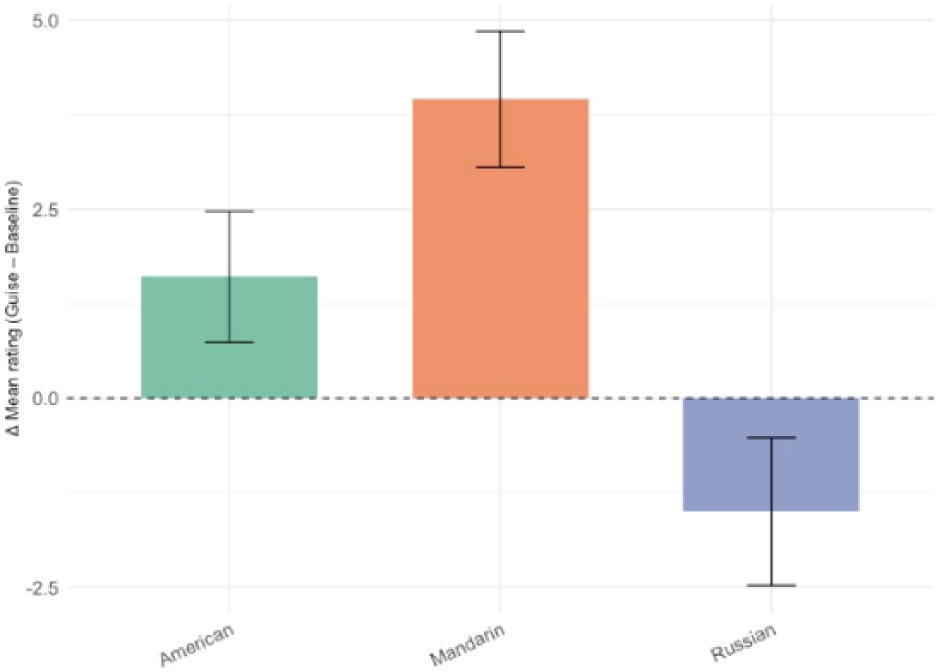


Figure 3.15: Russian–English bilinguals – alveolar voiced zone. Steps 1–5 are all collapsed.

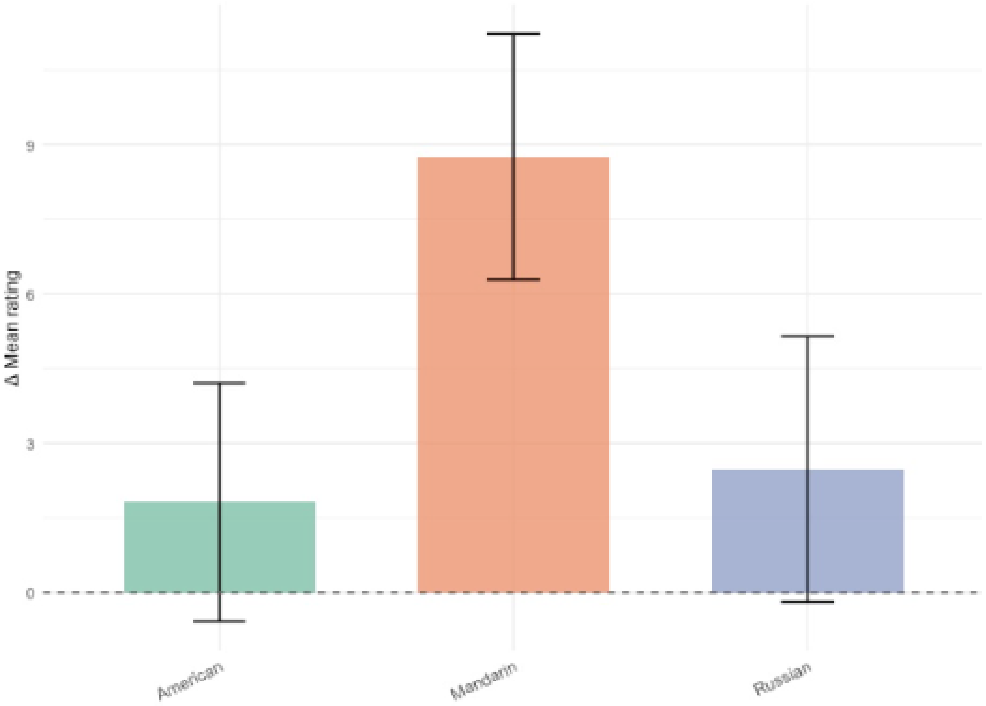


Figure 3.16: Russian–English bilinguals – velar voiced zone (step 4). Mandarin guise broadens perceptual ambiguity at the velar boundary.

Voiceless zones (Table 3.15; bottom) revealed bidirectional shifts. For bilabial stops, Mandarin guises decreased ratings overall ($\beta = -8.76$, $p < .001$) but increased them at Steps 7 and 9 (e.g., Step 7: $\beta = 7.10$, $p = 0.006$), indicating boundary-adjacent uncertainty in Figure 3.17. Velar stops showed American-guise suppression ($\beta = -7.74$, $p = 0.047$) and Russian-guise enhancement overall ($\beta = 14.43$, $p < .001$), but significant suppression at Steps 7–9 (e.g., Step 8: $\beta = -13.50$, $p = 0.021$), confirming ceiling distrust in Figure 3.18.

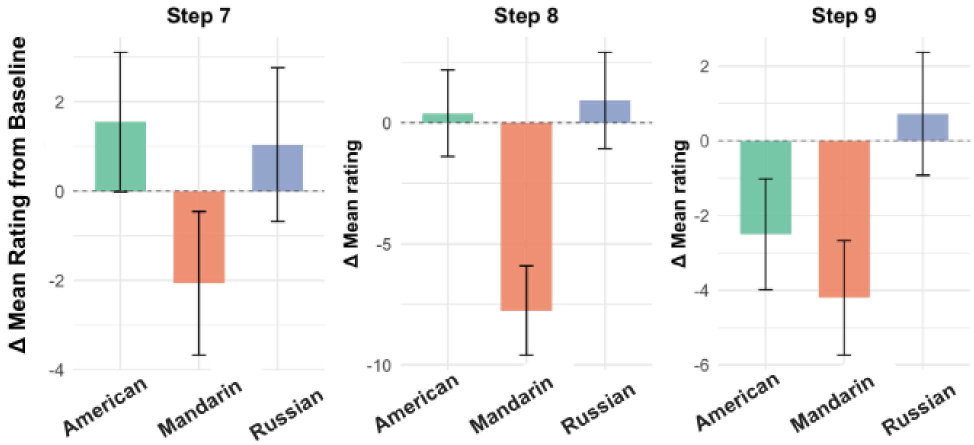


Figure 3.17: Russian–English bilinguals' bilabial voiceless zone (steps 7–9). Mandarin cues trigger uncertainty at Step 7, and this uncertainty increases as we move toward Step 9.

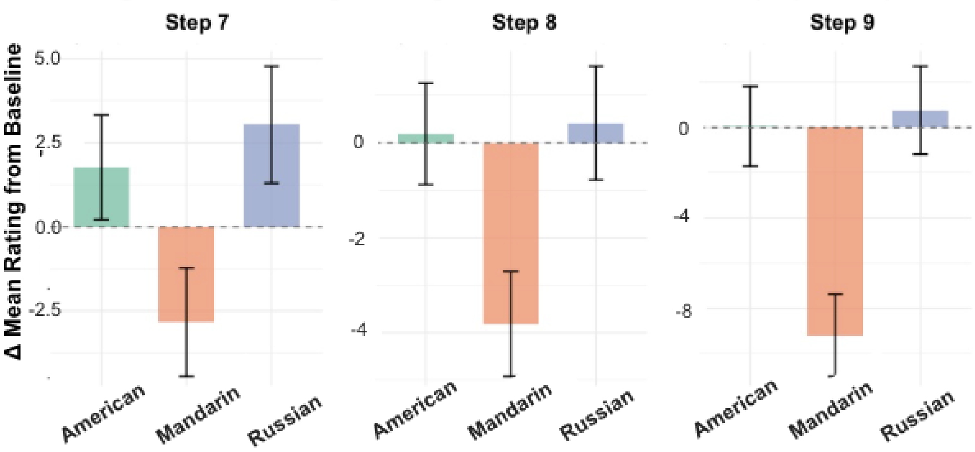


Figure 3.18: Russian–English bilinguals' velar voiceless zone (steps 7–9).

Ambiguous-Step Analysis. To resolve how social cues modulate core categorical perception—free from ceiling/floor contamination, I now isolate effects at precisely defined perceptual midpoints. Using baseline rating distributions (Figure 3.9), I identified six ambiguous steps where responses centered near 50%. See Table 3.16.

Table 3.16: Ambiguous step identification for both listener groups across PoAs.

Listener	PoA	Ambiguous Step	VOT (ms)
Mandarin	Bilabial	6	+40
Mandarin	Alveolar	6	+40
Mandarin	Velar	6	+50
Russian	Bilabial	5	-10
Russian	Alveolar	6	+20
Russian	Velar	6	+30

I fit a series of six linear mixed-effects models (one for each ambiguous-step subset defined in Table 3.16) to test whether speaker guise systematically shifts ratings at our most uncertainty-driven VOT midpoints. In each model, the dependent variable is the slider rating (`Value`), `Condition` (Baseline, American, Mandarin, Russian) enters as a fixed effect, and I include a random intercept for each participant (1 | `ID`) to account for repeated measures and individual-level baseline differences:

```
model <- lmer(Value ∼ Condition + (1 | ID), data = df_step)
```

The modeling revealed significant guise effects at the voicing boundary exclusively for Chinese listeners when confronted with the Russian guise at alveolar and velar positions (Table 3.17).

Table 3.17: Significant guise effects at voicing boundaries.

Listener	PoA	Step	Effect	β [95% CI]	t(Df)	p-value
Chinese	Alveolar	6	Russian – Baseline	13.08 [4.45, 21.71]	2.97 (238.2)	0.009
Chinese	Velar	6	Russian – Baseline	11.96 [3.44, 20.48]	2.75 (245.8)	0.018

The distributional patterns in Figures 3.19-3.20 illuminate how these statistical effects manifest on slider ratings. Under Russian guises, alveolar Step 6 responses shifted from baseline (M = 20.8) to 36.5 (SD = 37.2), expanding across the scale. Velar Step 6 showed parallel uncertainty amplification, with ratings doubling from baseline (M = 15.7) to 31.4 (SD = 35.6) under Russian guises.

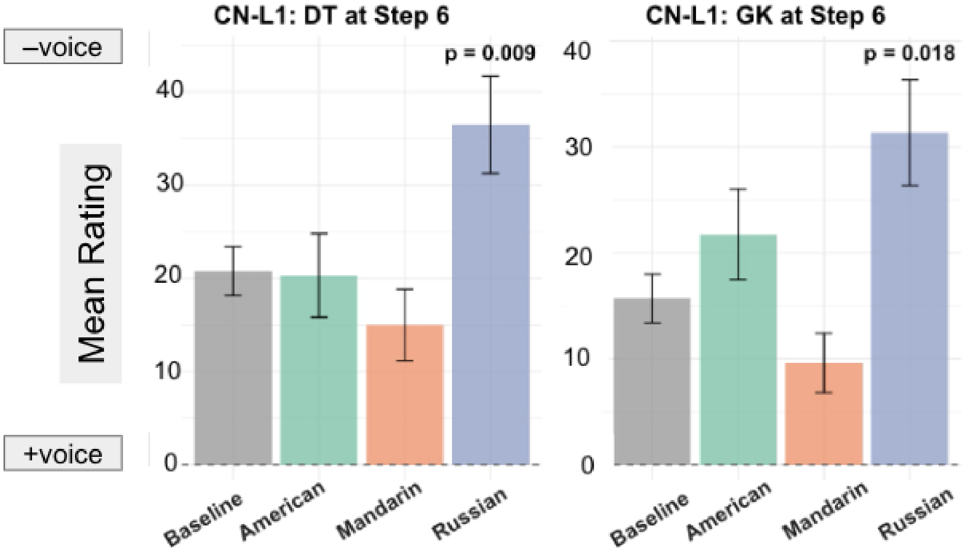


Figure 3.19: Significant boundary effects with baseline comparison. Includes baseline condition with gray bars; Russian listeners showed no significant effects.

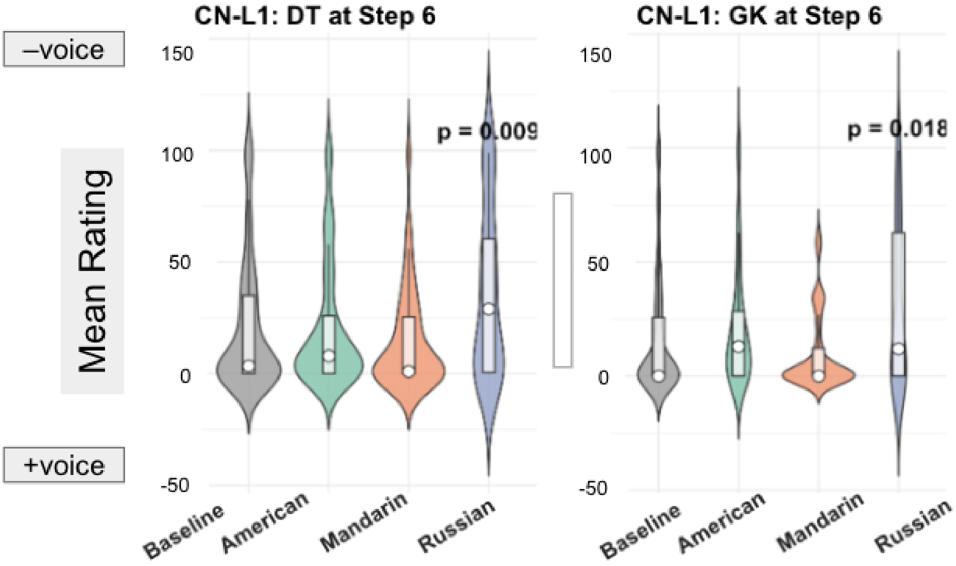


Figure 3.20: Significant boundary effects in violins. Gray violins show baseline distributions; *p*-values are shown for Russian vs. Baseline.

3.4 Results

This section reports findings for the three research questions (RQs) guiding our analysis: (**RQ1**) Do Russian–English and Mandarin–English bilinguals maintain L1-typical perceptual boundaries when categorizing voicing in English? (**RQ2**) Does talker identity (social guise) modulate voicing perception relative to baseline? (**RQ3**) Do these social-modulation effects align with phonological predictions derived from L1-based norms?

3.4.1 RQ1: Baseline VOT Boundaries

RQ1 asks whether Russian–English and Mandarin–English bilinguals maintain L1-typical boundaries when categorizing the voicing contrast in L2 English. Results show that, across all three places of articulation, both bilingual groups gravitate toward

English-like VOT categories, yet with notable asymmetries in their boundary placement. Mandarin bilinguals overshoot in the opposite direction, adopting shorter VOT thresholds than both L1 Mandarin and English norms. In contrast, Russian bilinguals converge toward English boundaries but do not fully reach the English mean, maintaining slightly shorter thresholds than English monolinguals.

At 30 ms VOT, for example, Russian listeners are more likely to categorize stops as voiceless (e.g., 41.2% for alveolar stops) compared to Mandarin listeners, who rarely accept 30 ms as voiceless ($<$ 8% for alveolars). This divergence reflects each group's L1 tuning: Russian's short-lag VOT predisposes earlier voiceless categorization, whereas Mandarin's long-lag VOT requires longer delays for voiceless perception. However, by 50 ms, both groups converge to near-categorical voiceless perception ($>$ 94% voiceless responses), suggesting convergence at the continuum's upper end.

These findings establish that both groups have adapted toward English VOT categories, but with Mandarin listeners "over-correcting" (undershooting) and Russian listeners partially converging toward English norms. These baseline patterns provide the critical reference framework for examining social modulation effects in RQ2.

3.4.2 RQ2: Social Modulation of Voicing Perception

RQ2 examines whether the presence of any talker identity (American, Mandarin, Russian guises) shifts listeners' voicing perception relative to baseline. The results show that talker identity does influence how bilingual listeners categorize the syllable voicing contrast, but it does so only in the narrow region of the continuum where the acoustic signal is genuinely ambiguous (e.g., Step 6).

For Mandarin-English bilinguals, the effect is highly focal: across all three places of articulation, a Russian guise increases voiceless ratings by approximately eight slider points at Step 6—the step that straddles their baseline boundary—while neither the American nor the Mandarin guise produces reliable shifts at any step of the continuum.

In contrast, Russian-English bilinguals exhibit a richer, bidirectional pattern. A Mandarin guise nudges the boundary forward at Step 5, then suppresses voiceless responses from Steps 7–9, while the Russian guise again raises voiceless judgments at Step 6. Although the direction of these shifts varies by guise, their locus is consistent across bilabial, alveolar, and velar stops, suggesting that the social effect is tied to the decision boundary itself rather than to fine articulatory details.

In other words, who listeners believe they are hearing matters most when the sound is ambiguous. Russian cues sharpen the perception of voicelessness for both groups, while Mandarin cues subtly reshape Russian listeners' boundaries but leave Mandarin listeners unaffected. However, when a sound is clearly voiced or voiceless, social cues do not play a significant role in how listeners judge the sounds.

3.4.3 RQ3: Directionality of Socially Modulated Shifts

Building on the baseline and social-modulation findings, RQ3 evaluates whether the observed guise effects align with L1-based directional predictions. To do so, I segmented the VOT continuum into voiced and voiceless zones tailored to each listener group and PoA, for example:

- Mandarin-English velar stops: voiced zone = Steps 1–6, voiceless zone = Steps 7–9.
- Russian-English bilabial stops: voiced zone = Steps 1–5, voiceless zone = Steps 6–9.

These empirically derived zones enabled testing whether guise-induced shifts aligned with phonological predictions.

It is hypothesized that Mandarin-English bilinguals would show a Mandarin guise decreasing voiceless perception (reflecting long-lag L1 norms) and an American guise increasing voicelessness (aligning with L2 expectations). Conversely, Russian-English bilinguals were expected to show Russian guise increasing voiceless perception (short-lag L1 norms) and American guise decreasing it. See Figures 3.21 and 3.22 for all directional effects.

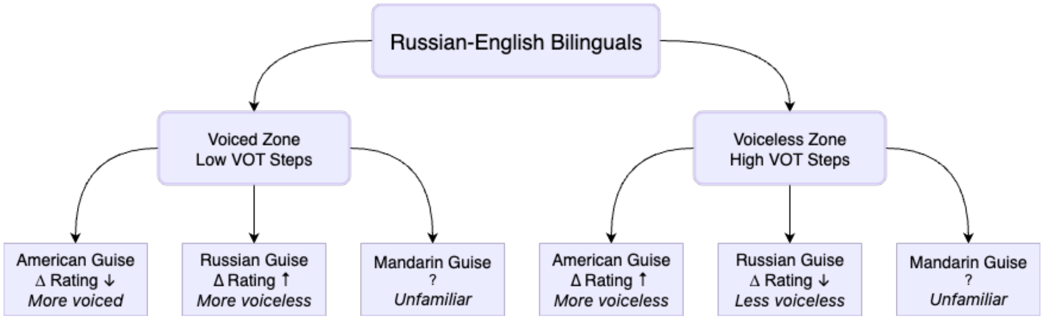


Figure 3.21: Russian–English bilinguals' expected guise effect on voicing rating.

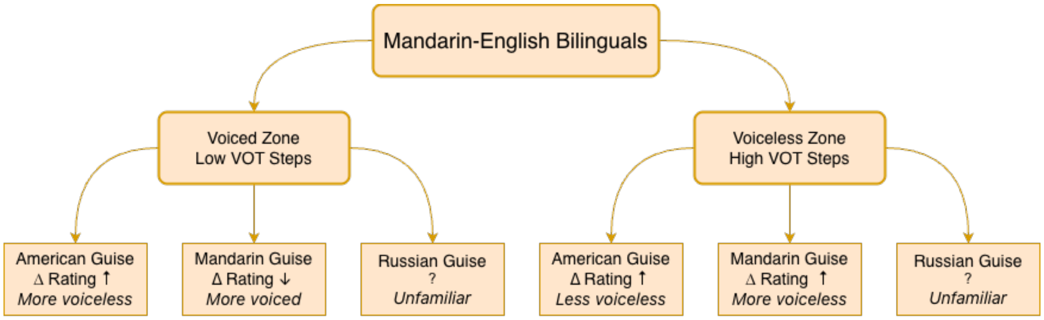


Figure 3.22: Mandarin–English bilinguals' expected guise effect on voicing rating.

Three key patterns emerge:

1. **Mandarin–English bilinguals**

    - **Voiced zones** showed the strongest effects across all three PoAs (bilabial: $\chi^2 = 11.47$, $p = 0.009$; alveolar: $\chi^2 = 12.85$, $p = 0.005$; velar: $\chi^2 = 18.75$, $p < .001$).
    - **Voiceless zones** were less consistently affected: bilabial ($\chi^2 = 9.66$, $p = 0.022$), alveolar showed a trend ($\chi^2 = 6.85$, $p = 0.077$), and velar effects were absent ($\chi^2 = 2.20$, $p = 0.961$).
    - **Russian guises** raised voiceless ratings around the categorical boundary, while American guises did not produce reliable shifts.

2. **Russian–English bilinguals**

    - **Voiceless zones** showed the strongest effects (bilabial: $\chi^2 = 26.95$, $p < .001$; alveolar: $\chi^2 = 12.54$, $p = 0.006$; velar: $\chi^2 = 9.84$, $p = 0.020$).
    - **Voiced zones** showed significance only for alveolar ($\chi^2 = 25.23$, $p < .001$) and velar ($\chi^2 = 36.28$, $p < .001$) stops.
    - **Russian guises** increased voiceless ratings (as predicted), while Mandarin guises exerted mixed or suppressive effects.

3. **PoA-specific sensitivity emerges**

    - **Velar stops** showed universally strong effects (all 4 tested zones were significant).
    - **Bilabial stops** exhibited asymmetric patterns: significant only in voiceless zones for Russians ($\chi^2 = 26.95$, $p < .001$) and in voiced zones for Mandarin ($\chi^2 = 11.47$, $p = 0.009$).

3.5 Discussion

Our findings provide three overarching insights into bilingual voicing perception. First, baseline boundaries (RQ1) reveal distinct L1-L2 negotiation strategies: Russian-English bilinguals partially shift toward English short-lag VOT norms, while Mandarin-English bilinguals undershoot both their L1 long-lag and L2 mid-lag expectations. This suggests a unique perceptual routine that does not align fully with either phonological system. [Either outcome, we support for the Unitary Language System (Volterra and Taeschner, 1978), whereby bilinguals' phonetic categorical perception lies intermediate between that of their L1 and L2, aligning with findings from Caramazza et al. (1973).

Second, social modulation (RQ2) emerges only when the acoustic signal is genuinely ambiguous, such as at category boundaries. Mandarin-English bilinguals show highly localized effects (at VOT step 6), while Russian-English bilinguals exhibit broader, bidirectional tuning to social cues.

Third, directional shifts (RQ3) largely align with phonological predictions[1], but also reveal novel mechanisms that extend existing models of speech perception. We attribute both expected and unexpected shifts to the following new mechanisms:

- **Uncertainty amplification**, where *unfamiliar* non-native guises widen perceptual ambiguity near category boundaries. For example, when Mandarin listeners encounter a Russian guise at the ambiguous Step 6, voiceless ratings nearly double. For velar stops, ratings increase from a baseline mean of 15.7 – indicative of a voiceless categorization) to 31.4 – towards uncertain, gradient

[1] Across all zones, 68% of significant shifts (12 of 17 effects) aligned with directional predictions.

response, reflecting possibly both a higher average and a broader distribution of responses (See discussion above Table 3.10)

- **Ceiling distrust**, where unfamiliar non-native guises prompt intermediate ambiguous responses even in clear acoustic zones when auditory stimuli come from continua ends. (See Figures 12–14, 18, 19; e.g. For Mandarin Listeners, a Russian guise paired with velar Step 9 reduced voiceless rating by 10.62 points).

3.5.1 Theoretical Implications

Mechanisms of Sociophonetic Interaction in Bilingual Speech Processing

Our results highlight two distinct yet interconnected pathways by which social identity interacts with bilingual speech perception, both modulated by acoustic ambiguity.

Pathway 1: Ambiguity-Driven Phonological Tuning. When the acoustic signal is ambiguous, bilingual listeners recruit social-indexed phonological priors to resolve uncertainty. This manifests as perceptual warping, where social cues shift category boundaries in L1-consistent directions. Importantly, this process is flexible and dynamic: both listener groups transiently adjusted their perception to unpracticed L3-like patterns, constructing provisional "accent prototypes" (e.g., "Russian-accented /ta/") within minutes of exposure, even without lexical context.

Pathway 2: Clarity-Triggered Confidence Dynamics. In unambiguous zones, social cues affect confidence rather than category assignment. Unfamiliar guises lead listeners to question otherwise clear acoustic signals, a response we interpret as "credibility filtering." This extends the Ideal Adapter framework (Kleinschmidt and Jaeger, 2015), suggesting that beyond acoustic reliability and statistical learning, perceived speaker credibility can constrain pre-lexical processing.

The Active Inference Framework. These pathways collectively position speech perception as active inference—a dynamic process where the brain continually remaps sounds using social context as Bayesian priors, a fundamental rethinking of where "social" ends and "acoustic" begins. For example, when encountering a Russian-guise /ta/, listeners do not merely classify the sound; they construct a context-specific prototype that integrates the following:

- **Acoustic input** (e.g., 40 ms VOT),
- **Social priors** (e.g., "Russian speakers produce shorter VOTs"), and
- **Credibility weighting** (e.g., "How much do I trust this talker or my own ability to discern their speech?")

This framework dissolves artificial boundaries between "acoustic" and "social" processing. Perceptual tuning (Pathway 1) and confidence calibration (Pathway 2) represent two phases of the same inferential continuum. In Phase 1, for ambiguous signals, social cues bias what we hear. In Phase 2, given clear signals, social cues bias how much we believe what we hear.

3.5.2 Limitations and Future Directions

Several methodological considerations warrant discussion. First, our within-subjects design—where each participant experienced all three social guises—was initially intended to maximize statistical power. However, empirical validation revealed significant block-order confounds, including systematic imbalances in guise distribution across blocks and carryover effects (particularly for Russian-English listeners). This necessitated restricting our primary analyses to `First_Social` blocks, effectively converting the design to a between-subjects comparison for social conditions. While this adaptation ensured analytical rigor against confounds (e.g., its potential to bias later social blocks) and is theoretically grounded in how social information accumulates over time, it reduced usable data volume by approximately two-thirds. Future studies could implement randomized between-subjects designs with single-guise exposure, or incorporate multi-session administration across blocks to preserve within-subjects advantages while mitigating carryover effects.

Second, our operationalization of perceptual boundaries, while theoretically grounded, relied on a $> 75\%$ VAS threshold rather than traditional two-alternative forced choice (2AFC) measures. This approach may lack precision in pinpointing exact categorical boundaries, potentially explaining why in baseline phase (RQ1), both listener groups exhibited identical boundaries (30 ms bilabial, 50 ms alveolar/velar) despite divergent L1 backgrounds. To enhance accuracy, future work should include a pilot 2AFC task to establish listener-specific boundaries before the main experiments.

Finally, unexamined individual differences (L2 proficiency, cultural identification) may modulate social cue weighting. Incorporating objective measures for L2 English proficiency and social identity questionnaires would allow modeling how bilingual experience predicts perceptual warping and distrust responses. In Chapter Five, we will continue to explore how individual language proficiency, attitudes, and socio-cognitive tendencies may correlate with the degree of phonological shifts captured in this chapter.

These limitations, while constraining generalizability, highlight opportunities for methodological innovation. Most critically, they underscore the need for paradigms that balance experimental control with ecological richness, preserving the social-phonetic negotiation we have documented while capturing its real-world dynamism. Future work addressing these constraints will deepen our understanding of how social meaning and acoustic signals co-construct perception in diverse linguistic ecosystems.

To conclude, speech perception is never passive reception, but an active construction site where social identity and acoustic signals are in constant negotiation. When sounds blur, social meaning tells us what to build. When sounds stand clear, social meaning tells us whether to trust the blueprint. In bilingual minds, this dialectic unfolds with virtuosic flexibility, revealing human perception as intrinsically social, all the way down to the syllable.

Chapter 4

BILINGUALS DIFFER IN WEIGHING SOCIAL AND ACOUSTIC CUES DURING ONLINE LEXICAL PROCESSING: EVIDENCE FROM EYE-TRACKING

4.1 Introduction

How do bilingual listeners resolve ambiguity in speech when both acoustic cues and social expectations are in play, and how does this process unfold in real-time? The previous chapter explored this question by examining how Russian-English and Mandarin-English bilinguals categorized voicing contrast in isolated syllable contexts using an explicit, continuous rating scale. However, speech perception in everyday contexts often involves interpreting meaningful lexical items or larger utterance chunks as they are heard, with perceptual decisions unfolding incrementally and automatically over time, and through subtle, continuous behavioral responses. The present chapter addresses how social expectations and acoustic variation interact in online word recognition, using a novel eye-tracking paradigm to capture real-time gaze behavior.

This experiment innovates on two fronts. First, it introduces a novel experimental paradigm to the study of bilingual speech perception by using eye movements to capture real-time categorization behavior. Second, it offers a new methodological framework for measuring gaze patterns in a scalable, cost-effective way. Unlike conventional eye-tracking studies that rely on expensive hardware and commercial software, I propose an open-access, technically flexible alternative for real-time gaze capture. Specifically, I (1) employ OpenFace 2.0 (Baltrušaitis et al., 2016), a freely available computer vision toolkit, to extract raw gaze angle data from simple webcam recordings of participants; (2) develop a custom technical pipeline to convert those gaze angles into interpretable screen-based spatial coordinates, mapped to our experimental layout; and (3) propose a set of behaviorally meaningful gaze encodings, such as direction of first look and gaze onset latency, which allow us to translate raw eye movement data into analyzable, valid measures of perceptual categorization. Together, these innovations create a robust infrastructure for studying bilingual speech perception without reliance on proprietary software or physical lab-based constraints.

This experiment tracks how the same bilingual groups (Russian-English and Mandarin-English bilinguals) process lexical items (e.g., bark vs. park) varying in VOT, under different social guise conditions. Using a self-paced, gaze-contingent interface, participants reveal what they heard by looking toward the side of the screen mapped to a voiced or voiceless interpretation. Eye movements are subconsciously controlled responses modulated by cognitive and attentional processes (Rayner, 2009). Their time-sensitive dynamics provide a powerful implicit measure of how early speech cues and social expectations are integrated in real time.

This implicitness is particularly important for studying socially modulated perception. Compared to explicit tasks like rating scales, gaze tracking is less susceptible to conscious control or social desirability biases, and can therefore reveal early perceptual biases that participants may otherwise suppress. Moreover, by using lexical items, this task allows us to investigate how bilingual listeners resolve ambiguity not just at the phonetic level, but at the level of lexical access, where expectations about both sound and meaning come into play.

To our knowledge, this is the first study to adapt the anticipatory eye movement (AEM) paradigm to bilingual speech perception while incorporating acoustic manipulations and social priming implemented through an accessible and flexible technical design. This work can serve as a new model for investigating how listeners use social and acoustic information during real-time language processing—especially in bilingual or second-language contexts where phonetic and social expectations often conflict.

In the sections that follow, I present the research questions guiding this experiment, followed by a detailed overview of the experimental design, technical implementation, and gaze data processing pipeline. The chapter concludes with results, statistical analyses, and a discussion of implications for bilingual speech processing and method development in sociophonetics.

4.1.1 Research Questions and Hypotheses

The current study explores how bilingual listeners categorize lexical **voicing contrasts** in real time, and whether their **gaze responses** reveal systematic modulation by acoustic and social expectations. In particular, I set out to address the following questions:

1. **RQ1: Can eye movements in a lexical decision context reliably capture categorical voicing perception?**

This question tests the viability of the AEM paradigm as a method for measuring categorical speech perception in bilinguals. It assesses whether participants' gaze behavior aligns with the expected direction (voiced = right; voiceless = left) when hearing acoustically unambiguous tokens in the absence of social guise. Our validation criterion is that at least 80% of participants must reliably make categorical gaze responses to endpoint stimuli in baseline trials. This ensures that AEM is capturing robust phonological categorization before additional manipulations are analyzed.

2. **RQ2: How, if at all, does social guise modulate listeners' gaze behavior when categorizing voicing contrasts in lexical items?**

This question investigates whether the perceived identity of the talker (e.g., native vs. non-native speaker) alters the way listeners integrate social and acoustic information during real-time speech processing. To address this, I analyze three distinct gaze-based measures:

1. **First Look Direction**: Did participants initially direct their gaze to the voiceless side?

2. **Onset Latency of Leftward Gaze**: How quickly did participants look toward the voiceless option?

3. **Total Duration of Leftward Gaze**: How long did participants sustain attention on the voiceless target?

I hypothesize the following outcomes to this question:

1. $\mathbf{H}_0$ (*null*): Social guise does not affect online processing behavior as measured through eye-gaze. Categorization patterns will remain consistent across guise conditions, with no systematic differences in gaze direction, latency, or duration.

2. $\mathbf{H}_1$ (*alt 1*): The **presence** of social guise affects online processing behavior. Participants will exhibit faster or slower gaze onset, relative to trials without a social guise baseline, when categorizing ambiguous stimuli under native or non-native conditions.

3. $\mathbf{H}_2$ (*alt 2*): The **type** of social guise (native vs. non-native) differentially affects perceptual categorization. That is, participants will show qualitatively different categorization patterns depending on the social identity of the talker.

4.2 Methods

4.2.1 Participants

All participants in this experiment had previously completed the VAS (Visual Analog Scale) study at least one week prior to participating in this study. This minimum one-week delay is implemented to reduce potential carryover effects from the VAS task, ensuring that participants would be less likely to recall specific acoustic details or response strategies. This spacing also served to minimize task fatigue and cognitive interference between sessions. Similar inter-session intervals have been recommended in prior research to reduce short-term memory effects and preserve the independence of experimental tasks (e.g., Freyaldenhoven et al. 2006; Golomb et al. 2007; Kong and Edwards 2011; Wade et al. 2021; Roark et al. 2023).

Participants had normal or corrected-to-normal vision in order to ensure accurate gaze tracking. A total of 51 individuals participated in the study: 27 native Russian speakers (9 male, 18 female) and 24 native Mandarin speakers (8 male, 16 female). Detailed demographic information will be reported in the results section.

4.2.2 Stimuli

Auditory stimuli

The auditory stimuli used in this experiment consisted of three English minimal pairs: BARK-PARK, DART–TART, and GUARD–CARD. Each pair was selected to represent a different place of articulation—bilabial, alveolar, and velar, respectively. All tokens were monosyllabic lexical items that are common and concrete in English, chosen to facilitate visual pairing and to minimize lexical frequency or semantic ambiguity effects.

First, I recorded original productions of these items from two male native speakers of American English using a Zoom H4N recorder in a sound-attenuated booth. Then, from these recordings, I constructed a 9-step voice onset time (VOT) continuum for each minimal pair, ranging from fully voiced (−70 ms) to fully voiceless (+90 ms). Using Praat, I interpolated VOT steps between endpoints while preserving natural spectral and temporal transitions. This yielded 27 unique auditory tokens (3 pairs × 9 steps), enabling fine-grained assessment of listeners' sensitivity to voicing contrasts. Stimuli were generated using a combination of Winn's (2020) Praat script and manual refinement when necessary, consistent with the procedures and techniques described in Chapter 3.

Social Guise Stimuli

This experiment introduced social expectations through two guise conditions: a native English speaker (henceforth, native) and a non-native English speaker whose first language matches the listener's (matching non-native). Since the listener groups consist of Russian-English and Mandarin-English bilinguals, three social guises are still created in total: American, Russian, and Chinese.

To operationalize these guises, three male actors of the corresponding cultural and linguistic backgrounds were recruited to deliver short self-introductions in their native languages (American English, Mandarin, or Russian). This design aimed to create realistic impressions that the auditory stimuli in the task were produced by the individual introduced in the video. Importantly, each actor was selected to visually represent a culturally prototypical exemplar of their respective linguistic background: an East Asian face for the Chinese guise, a European face for the Russian guise, and a Black face for the American guise. Because visual cues inevitably carry racialized meanings (Kutlu et al., 2022c,b; Kutlu, 2023), this research also considers how the speakers' racialized identities may interact with listeners' perception of speech and social guise. These choices were deliberate—particularly in contrasting the American and Russian guises—to maximize visual distinctiveness and promote perceptual separation between talkers. By maintaining roughly equidistant visual contrasts across conditions, the design aimed to enhance the activation of sociophonetic expectations linked to perceived speaker identity and to examine how those expectations might be influenced by racial cues embedded within the social guises.

All actors were instructed to deliver a 15–20 second video introduction featuring plausible but fictional biographical details, including name, hometown, and field of study as graduate students at Georgetown University. They wore formal clothing (e.g., suits or blazers) to minimize any perceived socioeconomic variation. See Figure 4.1 for the portrayed characters and their corresponding social guises. While actors themselves resided in different U.S. regions, the portrayed characters' uniform affiliation with Georgetown helped equalize perceived educational and occupational status.

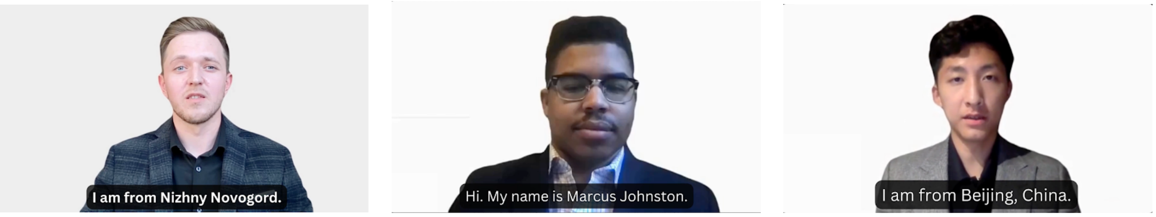

I am from Nizhny Novogord.
Hi. My name is Marcus Johnston.
I am from Beijing, China.

Figure 4.1: Actors representing the social guises used in AEM task. From left to right: Russian, American, and Chinese.

To ensure consistency, all introductions were scripted in advance, with actors rehearsing multiple times. The most natural and fluent take was selected for use. Each script included plausible biographical details designed to enhance authenticity. See Table 4.1 for character profiles and their scripted introductions. Subtitles in English were provided for all videos to ensure comprehension.

Table 4.1: Character profiles and scripted introductions.

Name	Guise	Introduction Script
Marcus Johnston	American	Hello, my name is Marcus Johnston. I'm from Baltimore, Maryland, and I'm a graduate student studying International Relations at Georgetown University.
Ilya	Russian	Здравствуйте, меня зовут Илья! Я из Нижнего Новгорода. Я аспирант, и я занимаюсь исследованием в области физики. *Translation: Hello! My name is Ilya. I'm from Nizhny Novgorod. I am a graduate student. I do research in the field of physics.*
Yang Yuchen	Chinese	大家好。我的名字叫杨雨辰。我来自于北京。我是一名经济学的研究生。 *Translation: Hello, everyone. My name is Yang Yuchen. I am from Beijing. I am now a graduate student studying Economics.*

Notably, in contrast to the VAS experiment (Chapter 3), participants in this task did not hear the non-native actors speak English. The design here was to investigate whether listeners, hearing only a speaker's native Russian or Mandarin, could infer how that speaker would sound in their L2 English. Importantly, none of the video introductions contained words with stop-initial consonants that were used in the task itself, thus eliminating any potential phonetic or lexical priming. This setup tests whether social identity alone is sufficient to trigger differences in listeners' categorization of the subsequent ambiguous auditory stimuli.

The AEM Simulation Creation

The AEM simulation paradigm was adapted from Kong and Edwards (2011), who used a Y-shaped anticipatory eye movement set up to examine listeners' perception of /da/–/ta/ contrast using visual targets of a dog and a taco. In the current version, custom simulations are created for each of the six target words: BARK, PARK, DART, TART, GUARD, CARD, using publicly available stock images to represent the referents.

All shapes and animations were constructed in PowerPoint, where I built the Y-shaped track (constructed from three rectangles) on a grey background (hex: #808080). The Y-channel, indicating the possible travel paths, was drawn in white (hex: #FDFDFD) to maximize contrast. Each simulation includes four visual layers: (1) a grey background, (2) a base Y-channel layer, (3) the moving object (image), and (4) a second top Y-channel layer, which can be adjusted in its level of transparency. Objects begin at the bottom center of the Y and travel either up-left (voiceless) or up-right (voiced), mimicking the auditory contrast structure. See Figure 4.2 below for an example of the traveling paths of objects during the DART-TART block.

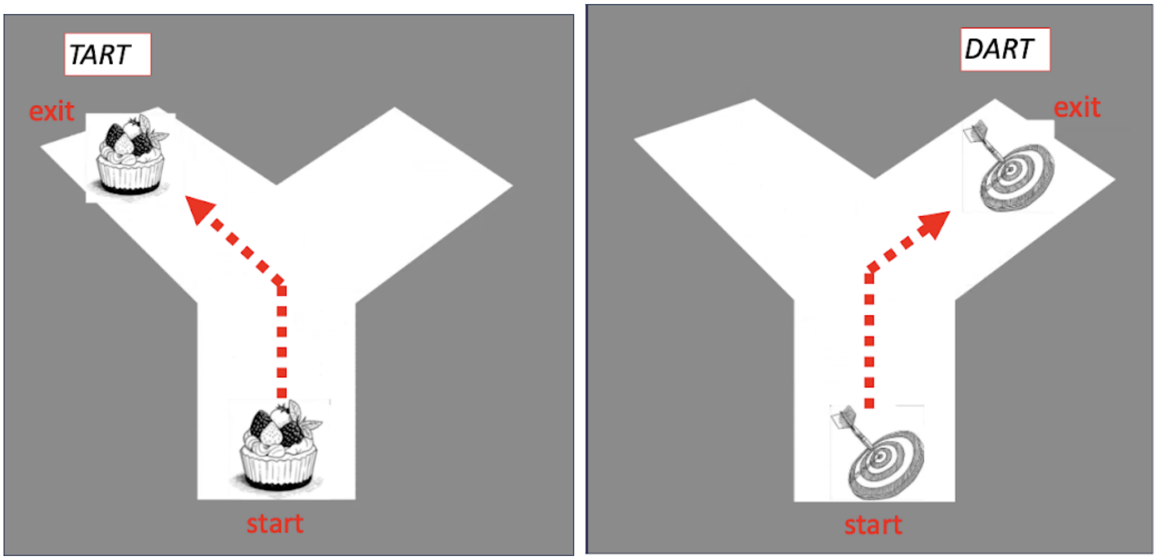


Figure 4.2: Example travel paths of the voiceless item (left) and the voiced item (right) in the DART–TART block.

PowerPoint's animation and layering tools enabled precise control over object movement and transparency, which was critical for the training phase. Transparency levels were manipulated by adjusting the top Y-channel layer, with 100 indicating full visibility of the traveling object and 5 indicating near invisibility.

I created three sets of simulation videos, each roughly 5 seconds long, varying in purpose and transparency:

1. **Training Phase (voice 1):** These simulations paired clearly voiced or voiceless speech tokens with three levels of Y-channel transparency:

   (a) 5 – The object is barely visible.

   (b) 60 – The object is faint but traceable.

   (c) 100 – The object is fully visible.

This gradual layering was designed to help participants associate voicing with spatial direction (voiceless = left, voiced = right) while relying less on visual cues and more on the auditory information over time.

2. **Baseline Phase (voice 1):** In these trials, the moving object was replaced with a traveling question mark, which disappears before reaching the fork. At the end of the trial, two static images (e.g., BARK vs. PARK) appear at the Y-channel endpoints (voiceless on the left, voiced on the right). This layout reminds participants of object positions, though they are expected to have made their gaze decision before the final images appear (see Figure 4.3). All videos here are paired up with a distinct VOT step across each word pair, therefore yielding a total of 27 unique video files (3 word pairs × 9 VOT step).

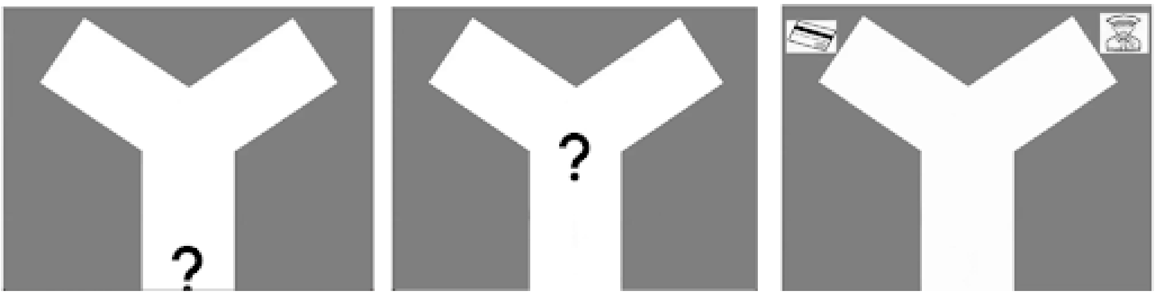


Figure 4.3: Example of a baseline trial in the GUARD–CARD block. The trial begins with a question mark emerging from the bottom center (left), moves toward the Y-channel fork (center), and ends with both target images appearing at the upper endpoints (right).

3. **Experimental Phase (voice 2 & voice 3):** Each experimental trial began with a brief video introduction of a talker (as described previously), followed by a simulation identical to the baseline phase but with new audio pairings. One voice represents the native guise, and another voice represents the non-native guise matching the listener's L1. These voices were perceptually distinct from one another and from the training voice (voice 1) to preserve the impression of distinct talkers across conditions. Importantly, the voice-to-guise pairing during this phase is not fixed. For each listener group, the native and non-native guises were randomly assigned to either voice 2 or voice 3. This counterbalancing was implemented using the Template function in PCIbex. The goal was to prevent any observed social effects from being attributable to specific acoustic properties of a single voice, thereby eliminating the potential confound of a voice-only effect.

4.2.3 Procedure

The Anticipatory Eye Movement (AEM) Task was fully built and administered via PCIbex and lasted approximately 25 to 30 minutes. The task was self-paced, as participants needed to press the spacebar at the end of each trial to proceed onto the next. All sessions took place in the Linguistics Lab using a 32-inch monitor, a Mac laptop, a high-frame-rate webcam, noise-canceling headphones, a mouse, and a keyboard. The monitor was placed at the back of the desk, and behind a webcam on a tripod which was positioned roughly 1.5 feet away from the participant and angled slightly upward to capture only their face. The laptop was set to the participant's left so they could press the spacebar with their left hand while maintaining gaze focus on the screen.

Recording Setup

OBS Studio is used to simultaneously record two screens and a webcam feed (See Figure 4.4 for an example of an OBS captured video). The OBS setup included three sources:

1. Screen Capture of the experiment, showing up at the bottom-right of the studio.
2. Video Capture Device for the webcam, placed in the bottom-left corner.
3. Browser Source showing a live timestamp (from `currentmillis.com`) above the participant's face for synchronization. This external timestamp allowed us to later align participants' gaze movements, as extracted from webcam video, with the onset time of each trial in the PCIbex experiment. Because I did not use PCIbex's built-in eye-tracking function, this step was essential for validating and pairing gaze data with the experimental timeline.

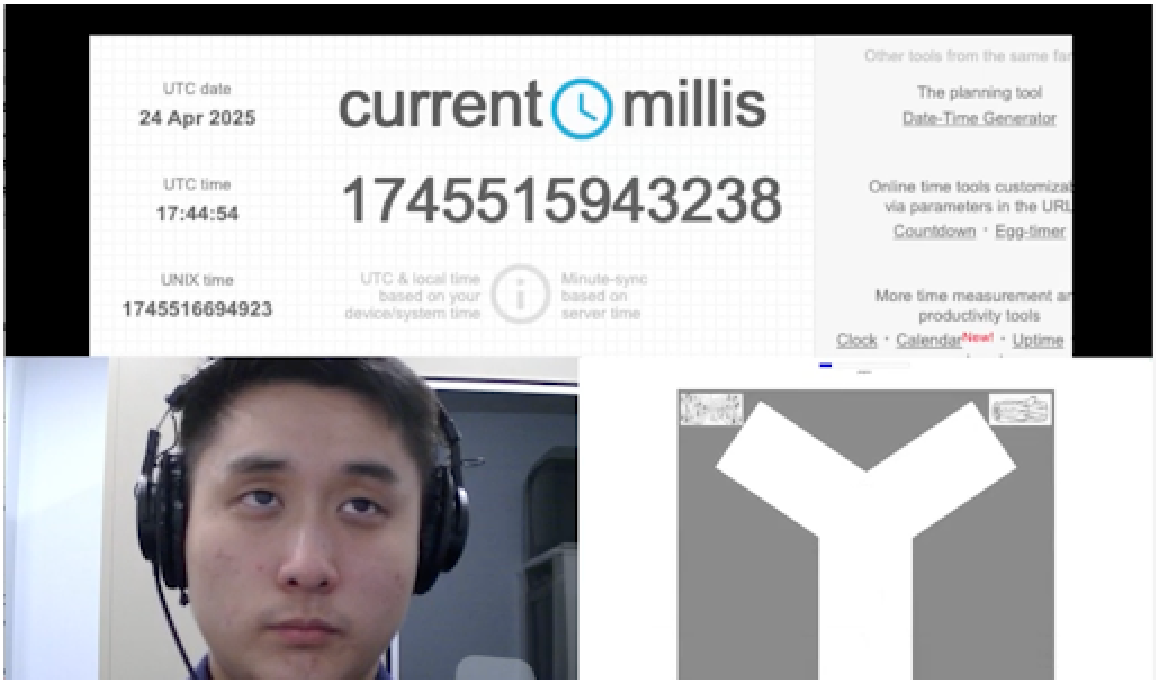

UTC date
24 Apr 2025
UTC time
17:44:54
UNIX time
1745516694923
current millis
1745515943238
UTC & local time based on your device/system time
Minute-sync based on server time
Other tools from the same fa
The planning tool
Date-Time Generator
Online time tools customiza via parameters in the URL
Countdown · Egg-timer
More time measurement ar productivity tools
Clock · Calendar New! · Uptime

Figure 4.4: Example video frame exported from OBS during an experimental session.

Before beginning the experiment, I calibrated the setup by adjusting the camera angle and confirming the participant's comfort and ability to view the screen clearly. Participants were then instructed, both verbally and via the text instructions on screen, to follow the moving objects with their eyes only and to minimize head and shoulder movement during the task.

I conducted a test recording trial to confirm that screen capture, webcam alignment, and audio-visual synchronization were all functioning correctly. During this stage, participants are instructed to follow the cursor around the screen using only their eyes. The test video was reviewed for alignment, facial visibility, and timestamp readability. All instructions were provided in the participants' dominant language.

The experiment consisted of two phases: (1) an integrated training-baseline phase and (2) the experimental phase. Each phase lasts 12-15 minutes, with a short break in between. During the break, participants were encouraged to stretch, leave the lab, or get water during the break. All trials were built and presented using PCIbex, and participants wore noise-canceling headphones throughout.

Training and Baseline Phase

The training and baseline phases were structured by place of articulation (POA) and presented in **blocked order**: first **BARK–PARK**, then **DART–TART**, and finally **GUARD–CARD**. The experiment did not interleave across POAs; each block was completed in full before moving to the next.

Each POA block began with a **training phase** that familiarized participants with the word-object mappings and the movement patterns associated with voicing. Participants first saw the visual object corresponding to the **voiced** word travel along the Y-channel three times (top-right exit), followed by the **voiceless** word object three times (top-left exit). Next, they saw an interleaved series of six trials with decreasing object visibility: the voiced and voiceless words alternated, first at 60% transparency and then at 5%, with each word presented once at each level. This resulted in 10 training trials per training block. The purpose of this gradual transparency reduction was to train participants to associate the auditory cue with directionality (voiced = right, voiceless = left) and ultimately rely less on vivid visual information.

Immediately following training, participants completed the **baseline perception task** for the same POA. Each baseline trial used the same Y-shaped simulation, but now featured a question mark traveling upward, disappearing before the Y-fork. At trial end, two static images (e.g., BARK and PARK) appeared at the endpoints of the Y-channel. The auditory stimulus followed a 9-step VOT continuum, with each step repeated 3 times, totaling 27 **trials** per POA block. This training, followed by a baseline procedure, was repeated for each POA. After completing all three blocks, participants were given a short break before continuing to the experimental phase.

Experimental Phase

In this phase, participants encountered two speaker guises: a native English speaker and a non-native speaker matching the participant's L1. Each guise began with a short introductory video, followed by a simple comprehension question to ensure listeners' engagement. The experimental trials were structurally identical to the baseline trials but paired with Voice 2 (native guise) or Voice 3 (non-native guise). The voice-to-guise assignment was randomized and counterbalanced across listeners. This was done to prevent potential voice-specific effects from confounding social interpretation effects. See Figure 4.5 for the entire workflow of AEM.

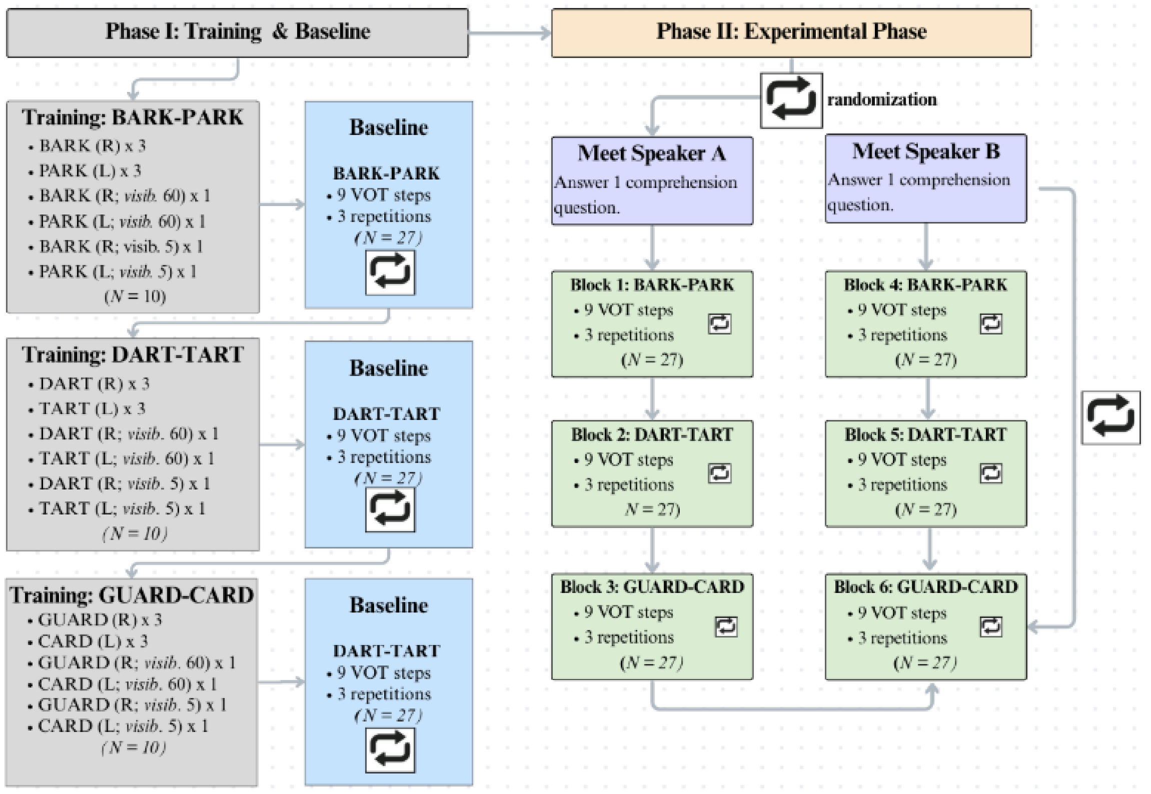


Figure 4.5: Workflow and block structure of the AEM task.

Supplementary Individual Differences Assessments

Participants completed a battery of tests assessing language background and cognitive capacities, including the LEAP-Q, a working memory task, and the Autism Quotient. Results from the supplementary questionnaires, and their correlations with VAS and AEM performance are reported in Chapter 5.

4.3 Anticipatory Eye Movement Data Processing Pipeline

This section details the full processing pipeline used to transform raw audiovisual recordings and experimental outputs into structured gaze features suitable for statistical modeling. The goal was to extract interpretable, trial-aligned indicators of participants' anticipatory eye movements (AEM), enabling comparisons across conditions and listener groups. Below, I describe how OpenFace-derived gaze vectors, PCIbex trial logs, and externally embedded UNIX timestamps are integrated to generate a clean, feature-rich dataset.

4.3.1 Raw Data Sources and Structure

PCIbex Experimental Logs

Each experimental session generated structured trial logs via PCIbex, with data including participant ID, group, condition (social guise), phonological contrast (e.g., BARK–PARK), video file, VOT step, behavioral response, and a timestamp for each trial onset. These timestamps enabled approximate stimulus timing, but desynchronized slightly from actual video playback due to machine-level timing differences.

Video Recording and External Unix Timestamps. To establish a common timing reference, a live UNIX timestamp (`unixtimestamp.com`) is embedded in each screen recording using OBS Studio. This visual timestamp served as a bridge between

PCIbex's internal timing and the external webcam video, allowing frame-accurate alignment between stimulus presentation and facial/gaze behavior.

OpenFace Output. To extract frame-by-frame gaze information from participants' webcam videos, each recording is processed via OpenFace 2.0, an open-source toolkit for facial behavior analysis. OpenFace applies a deep learning–based model to detect and track 68 facial landmarks across frames, including points around the eyes, nose, jawline, and mouth. It uses these landmarks to estimate both head pose and eye gaze direction.

The videos were originally captured at 260 Hz but were processed at 60 Hz due to OBS compression settings. From each frame, I extracted the following OpenFace outputs:

1. **Eye gaze direction vectors** (`gaze_0_x/y/z, gaze_1_x/y/z`): 3D unit vectors representing the direction of gaze from each eye.

2. **Head translation** (`pose_Tx, pose_Ty, pose_Tz`): position of the participant's head relative to the camera, describing horizontal/vertical displacement and distance from the lens.

3. **Head rotation** (`pose_Rx, pose_Ry, pose_Rz`): orientation of the head in space, expressed as rotations around the camera pitch, yaw, roll axes.

These raw features formed the input for subsequent geometric transformations and normalization steps, described in later sections.

4.3.2 Temporal Alignment of Trials and Gaze Data

To analyze participants' gaze behavior relative to each experimental trial, it is crucial to align the timing of PCIbex trial onsets with the frame-level outputs from OpenFace.

This alignment step required integrating information from two sources: the PCIbex-generated timestamp for each trial, and the real-time UNIX timestamp embedded in each screen recording.

I began by manually reading the embedded UNIX time visible in the OBS videos. For each experimental session, I recorded the screen time shown at the moment the first trial appeared, as well as the time corresponding to a known later trial (typically trial 2). These reference points allowed me to estimate the offset between PCIbex-reported trial start times and the actual system time shown in the video. This offset was necessary because PCIbex's timestamps, while internally consistent, did not always match the local time on the recording machine. On average, I observed a discrepancy of up to ±250 milliseconds between the PCIbex timestamp and the true video onset.

Using this time offset, I created a mapping between each PCIbex trial and a specific segment of the video recording. I then extracted the corresponding video frames for each trial and processed only those segments through OpenFace. This ensured that each trial was represented by a precise, temporally aligned sequence of facial and gaze data.

By anchoring the start of each gaze segment to a real-time UNIX timestamp, it was possible to reconstruct a continuous timeline of where participants were looking during each trial. This alignment step served as the foundation for all subsequent segmentation, smoothing, and feature extraction.

4.3.3 Converting Gaze to Screen Location

OpenFace does not provide screen-referenced gaze coordinates. To address this problem, I implemented a geometric projection method via a custom function:

estimate_screen_gaze_locations_by_head_pose(). This function performs the following:

1. Averages the 3D gaze direction vectors from both eyes (gaze_0_x/y/z, gaze_1_x/y/z).
2. Shoots a ray from the participant's estimated head center (pose_Tx, pose_Ty, pose_Tz) along the gaze direction.
3. Calculates the (x, y) intersection of this ray with a virtual plane approximating the screen.

The result was a **continuous estimate of gaze location** on the screen in horizontal (x) and vertical (y) dimensions, expressed in millimeters. I refer to this output as gaze_screen_x and gaze_screen_y.

To improve the stability and interpretability of these projections, I applied a two-stage normalization procedure:

1. **Head position normalization** (pre-gaze extraction): Before projecting gaze rays, I normalized head position by subtracting the median horizontal and vertical head translation (pose_Tx, pose_Ty) over the trial. This step stabilized the perceived location of the head within the camera frame, reducing session-level bias caused by inconsistent video cropping or seating variation.
2. **Screen gaze normalization** (post-projection): After screen gaze coordinates were extracted, I subtracted the trial-level median screen gaze (x, y) which defines a neutral or "center" gaze baseline. This normalization allowed us to classify gaze direction as left, center, or right relative to the participant's own head and seating position.

This approach was validated by having the author trace a rectangle with their eyes while allowing head movement, making the scenario more challenging than our lab-recorded sessions. The resulting gaze trajectory formed a recognizable shape, confirming that the pipeline recovered spatial gaze information reliably (See Figure E.1 in Appendix E.

4.3.4 Directional Classification and Segmentation

Once normalized screen gaze values were computed, each frame in a trial was classified as **left, center**, or **right** based on `gaze_screen_x`. Thresholds were applied as follows:

- Below -0.2 mm = `left`
- Between -0.2 and $+0.2$ mm = `center`
- Above $+0.2$ mm = `right`

This classified sequence was segmented into blocks of consistent gaze direction. To reduce jitter and improve stability, I applied a **median filter** over a 10-frame window ($\approx$ 167 ms at 60 Hz), replacing each frame's classification with the majority value in its window. This preserved genuine shifts in gaze while eliminating spurious frame-level flickers.

Each trial was now represented by a sequence of stable, directionally labeled gaze segments suitable for analysis.

4.3.5 Segment Correction and Gaze Encoding

Once directional segments were identified, I applied a final correction step to ensure segment boundaries reflected the true start and end of consistent gaze. Because smoothing via the median filter can shrink segments, delaying their onset and prematurely cutting them off, I used an expansion procedure that referenced the original (unsmoothed) classification. This restored any early or late frames that had been mistakenly excluded during smoothing.

From these corrected segments, I derived a set of trial-level gaze segment features to summarize the segmented gaze trajectory in interpretable terms. For each trial, I recorded the following information:

- Onset and duration of the first leftward and rightward looks segments, in milliseconds:
  (`first_looks_left_time_ms`, `first_looks_left_duration_ms`);
  (`first_looks_right_time_ms`, `first_looks_right_duration_ms`)
- Onset and duration of the last leftward and rightward looks segments, in milliseconds:
  (`last_looks_left_time_ms`, `last_looks_left_duration_ms`);
  (`last_looks_right_time_ms`, `last_looks_right_duration_ms`)
- Binary indicators for:
  - Whether the trial's first off-center look segment was leftward:
    (`first_looked_left`)
  - Whether the trial's last off-center look segment was leftward:
    (`last_looked_left`)
  - Whether the first-to-last segment gaze direction changed during the trial:
    (`fixation_flipped_from_first_to_last`)

An example of a segmented gaze trajectory with annotated features is shown in Figure 4.6. This figure illustrates how directional gaze segments are identified, labeled (e.g., first vs. last), and color-coded, providing the structural basis for the feature set used in later analysis.

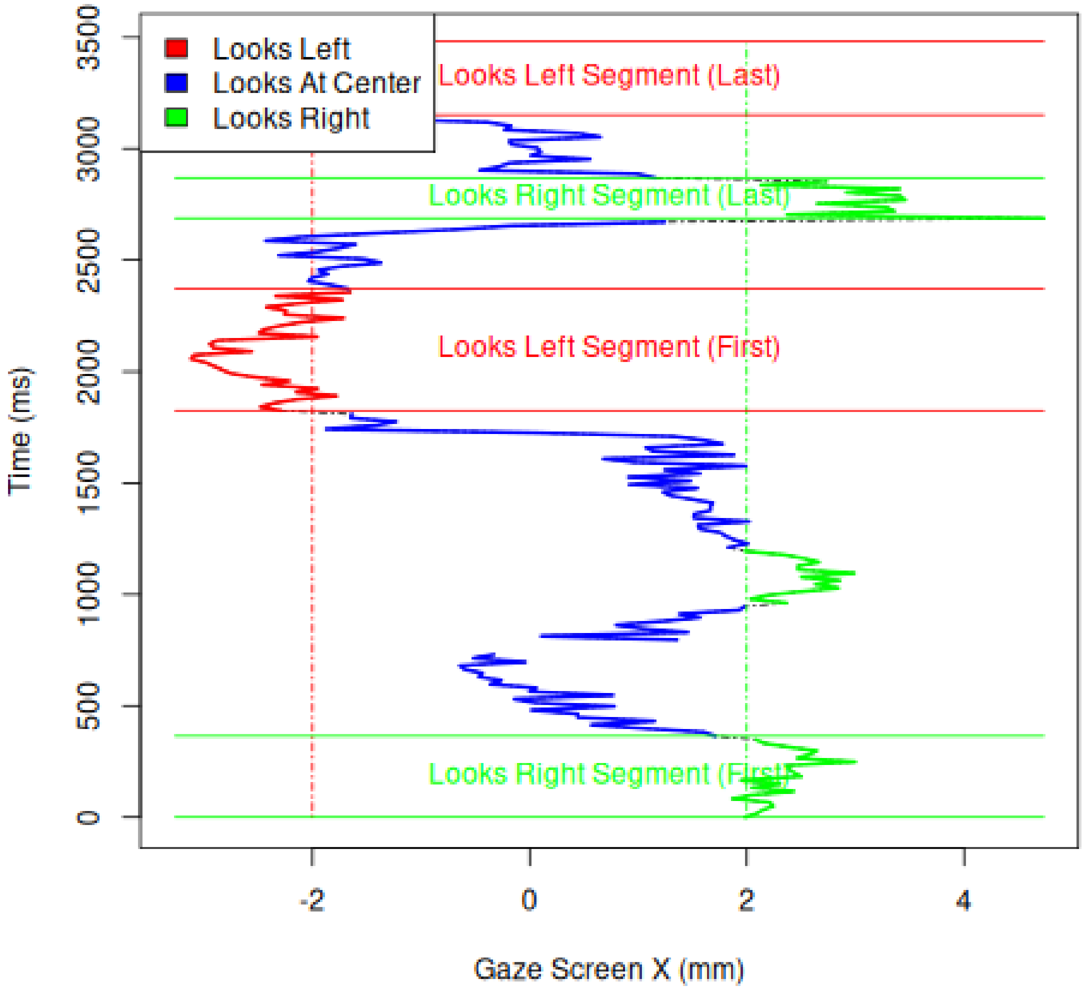


Figure 4.6: Example gaze segment encoding for a single trial.

The following gaze segment feature values are extracted for the scenario from the figure:

- `first_looks_left_time_ms` = 1824.6
- `first_looks_left_duration_ms` = 547.4

- `last_looks_left_time_ms` = 3151.6

- `last_looks_left_duration_ms` = 331.7

- `first_looks_right_time_ms` = 0

- `first_looks_right_duration_ms` = 364.9

- `last_looks_right_time_ms` = 2687.1

- `last_looks_right_duration_ms` = 182.5

- `first_looked_left` = FALSE

- `last_looked_left` = TRUE

- `fixation_flipped_from_first_to_last` = TRUE

These features enabled direct comparisons across trials and participants and were designed to be both interpretable and statistically tractable.

In parallel, I calculated classification-based duration features, based on the raw frame-level gaze direction labels:

- `total_cropped_looks_left_duration_ms`

- `total_cropped_looks_right_duration_ms`

- `total_cropped_looks_at_center_duration_ms`

These features encode the total times spent looking left or right across all frames in the trial, regardless of segmentation, in milliseconds. The following gaze duration feature values are obtained for the scenario from Figure 4.6:

- `total_cropped_looks_left_duration_ms` = 862.5

- `total_cropped_looks_right_duration_ms` = 763.0
- `total_cropped_looks_at_center_duration_ms` = 1808.0

All features were extracted from the first 3,500 ms of each trial. Trials lasted up to 4 seconds, but the final 500 ms are excluded to avoid capturing post-response or off-task behavior. This cropping ensured our metrics reflected real-time perceptual processing.

For statistical analysis, I focused on three key features: `first_looked_left` (binary), `first_looks_left_time_ms` (onset), `total_cropped_looks_left_duration_ms`.

4.3.6 Regional Threshold Tuning and Final Selection

To determine the most reliable and interpretable version of our dataset, I tested combinations of two pipeline parameters:

1. **Trial-level centering:** Whether the `gaze_screen_x` trajectory was median-centered per trial (as described in Section 4.3.3).
2. **Region threshold width:** How wide the neutral center region was defined for classification. I tested thresholds of ±2 **mm**, ±3.5 **mm**, and ±5 **mm**.

Each combination produced a variant of the dataset, which I evaluated using two criteria:

1. **Visual inspection of gaze classification across VOT steps.**
2. **Statistical alignment with expected perceptual patterns.**

Wider thresholds yielded more trials classified as beginning from the "center," increasing retention but also reducing sensitivity to directional bias. Centering removed persistent head bias but introduced greater variability across participants if done improperly.

Ultimately, I selected the version with **trial-level centering** and a ±2 **mm center region**. This dataset offered the best trade-off between interpretability, participant retention, and model performance. It served as the final input for all subsequent statistical analyses.

4.4 Statistical Analysis of Gaze Behavior

This section outlines the statistical approach used to address our research questions about gaze behavior during voicing contrast categorization. I analyze data from a bilingual participant sample using generalized and linear mixed-effects models to test how gaze responses are shaped by acoustic input and social guise.

I begin in Section 4.4.1 by addressing RQ1, which evaluates whether the Anticipatory Eye Movement (AEM) paradigm captures reliable, categorical perception of voicing contrasts in baseline trials. The remaining sections address RQ2, asking whether, and how, social guise influences real-time categorization. We examine three gaze-based measures:

- **First look direction** (Section 4.4.2) captures early categorization decisions;
- **Onset latency** (Section 4.4.3) measures how quickly listeners commit to a voiceless interpretation;
- **Total leftward gaze duration** (Section 4.4.4) indexes the strength and persistence of that perceptual commitment.

For each measure, we fit separate models by listener `group` × `POA` subset, using `vot_step` and `condition` variables as fixed effects and add random intercepts for the participant variable. Analyses are conducted both across the full VOT continuum and at ambiguous VOT steps, identified via predicted categorization probabilities or fixed at Step 6 for time-based measures. This modeling framework allows us to test whether social guise shapes not only what listeners categorize, but when and how strongly they do so in real time

4.4.1 Validating the AEM Paradigm for Speech Perception (RQ1)

Our first analytical goal was to establish whether anticipatory eye movements (AEM) could reliably capture categorical voicing perception in a lexical decision context. Given that the AEM paradigm is relatively novel and only minimally explored in speech perception research (Kong and Edwards, 2011), it was essential to first demonstrate its basic functionality and robustness within our bilingual listener population. To do so, we undertook a careful preprocessing and validation procedure designed both to assess the paradigm's effectiveness and to prepare a high-quality dataset for subsequent analyses.

Preprocessing of Sessions

We tested 51 bilingual participants (24 Chinese, 27 Russian), each completing two sessions: one baseline (acoustic cues only) and one experimental (with social guise). From 102 total recordings, we excluded sessions with major technical issues, such as missing PCIbex logs, poor-quality video, or insufficient facial landmark tracking by OpenFace. This processing led to the retention of 93 usable sessions, representing a high technical retention rate of 93.5%.

Accuracy-based Validation. Next, we implemented an accuracy-based validation check to assess whether the retained sessions indeed reflected reliable categorical voicing perception. We applied this check to **baseline sessions only**, which served as the most direct test of whether participants could map clear acoustic signals to the correct visual side. If participants failed to respond categorically in this condition, it would suggest either a lack of engagement or a breakdown in the AEM paradigm itself, thus making any interpretation of social modulation effects in the experimental blocks unreliable.

To evaluate this effect, we examined participants' gaze responses on the endpoint steps of the VOT continuum:

- On Step 1 trials (unambiguously voiced), the correct response was **NOT** to look left (`first_looked_left == FALSE`).
- On Step 9 trials (unambiguously voiceless), the correct response was to look left (`first_looked_left == TRUE`).

For each baseline session, we calculated the proportion of trials meeting these expected categorical gaze responses and applied a conservative accuracy threshold of 70%. Sessions that did not meet this criterion were excluded from further analysis. This step served two key functions: first, it explicitly validated the effectiveness of the AEM paradigm in capturing participants' categorical perception responses (addressing RQ1 directly); and second, it ensured that the final analytic sample consisted solely of participants who provided reliable and interpretable gaze data for subsequent analyses (RQ2 and RQ3).

After filtering, 87 sessions remained (42 baseline, 45 experimental), yielding an accuracy-based retention rate of 93.55% from 93 usable sessions and 85.29% from all 102 recorded sessions. This high rate of compliance — both across listener groups and experimental phases — supports the viability of AEM as a method for capturing real-time phonological categorization.

Figure 4.7 shows a breakdown of session counts by listener group and gender before and after applying the accuracy filter. Figure 4.8 illustrates accuracy on endpoint trials across all three conditions: baseline, native, and non-native. Notice that only in the baseline condition do combined responses to Step 1 and Step 9 consistently exceed the 70% threshold (marked by the red dashed line). In experimental blocks, accuracy was more variable, as expected, since social guise was hypothesized to bias perception. Sessions with subthreshold accuracy in native or non-native conditions (e.g., Participant `021F_CN_Sh_AEM-social`) were still retained, because the accuracy filter applied only to baseline blocks.

Taken together, these filtering steps and high retention rate confirm that our novel AEM paradigm elicits categorically meaningful gaze responses and can provide a solid foundation for analyzing more nuanced social modulation effects in subsequent sections.

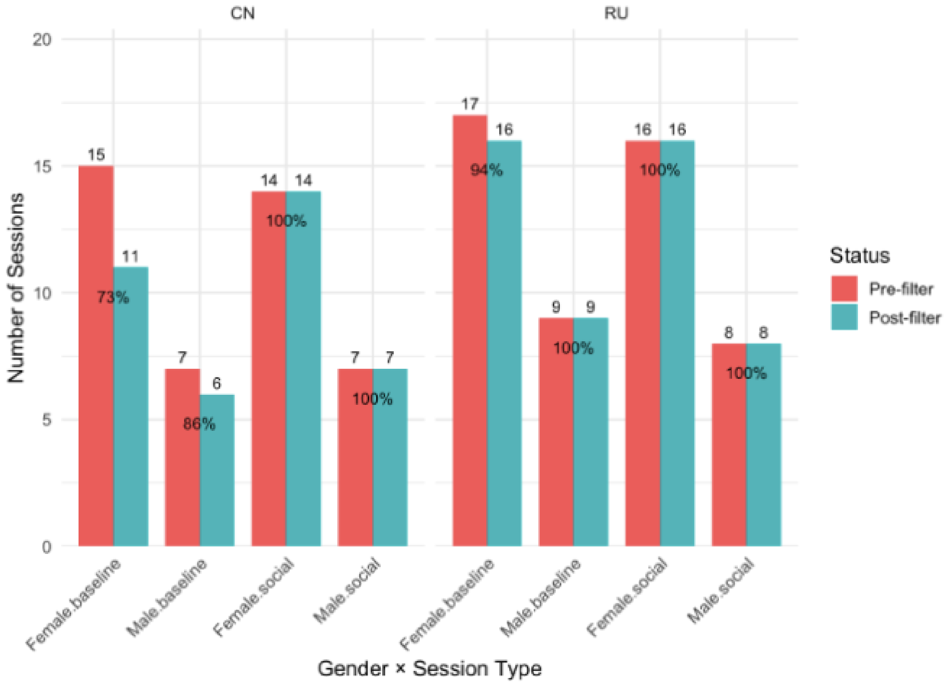


Figure 4.7: Participant counts and retention rate by listener group: pre- vs. post-filtering.

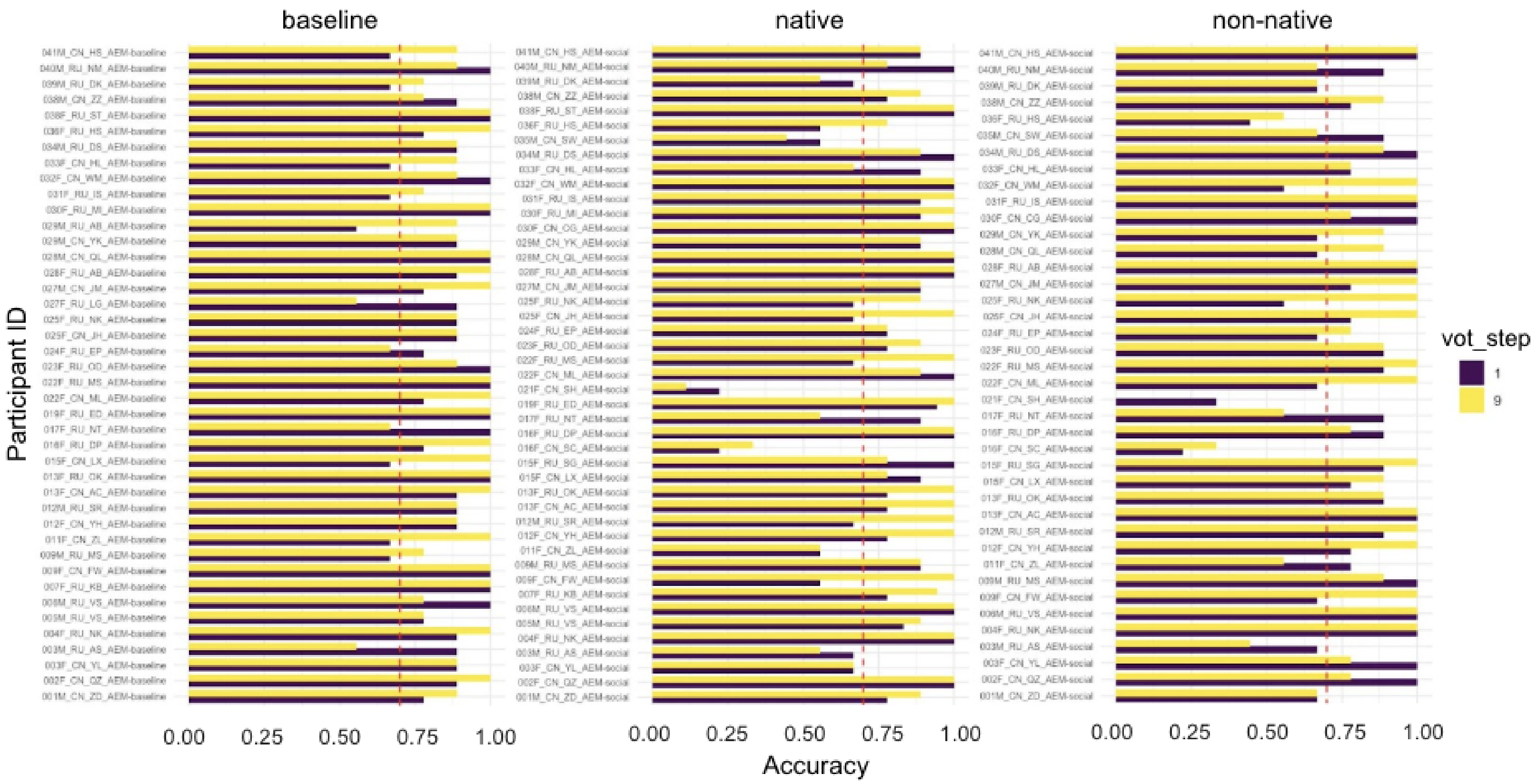


Figure 4.8: Accuracy rate by participant and VOT step across baseline and experimental conditions.

4.4.2 Social Modulation of Gaze Behavior: First Look Direction (RQ2)

This section examines whether social guise biases the initial perceptual categorization of voicing contrasts in real time, as reflected in the directionality of the first off-center gaze. Our binary dependent variable, `first_looked_left`, serves as a proxy for voiceless categorization. This metric captures the initial categorical decision, offering a discrete index of how social expectations may sway early perceptual commitment across VOT continua. To visualize group-level patterns in initial categorization, Figure 4.9 plots the proportion of first looks to the left (voiceless) across all VOT steps, by condition and listener group. This provides a descriptive overview of how gaze responses shift along the continuum and under different social guises.

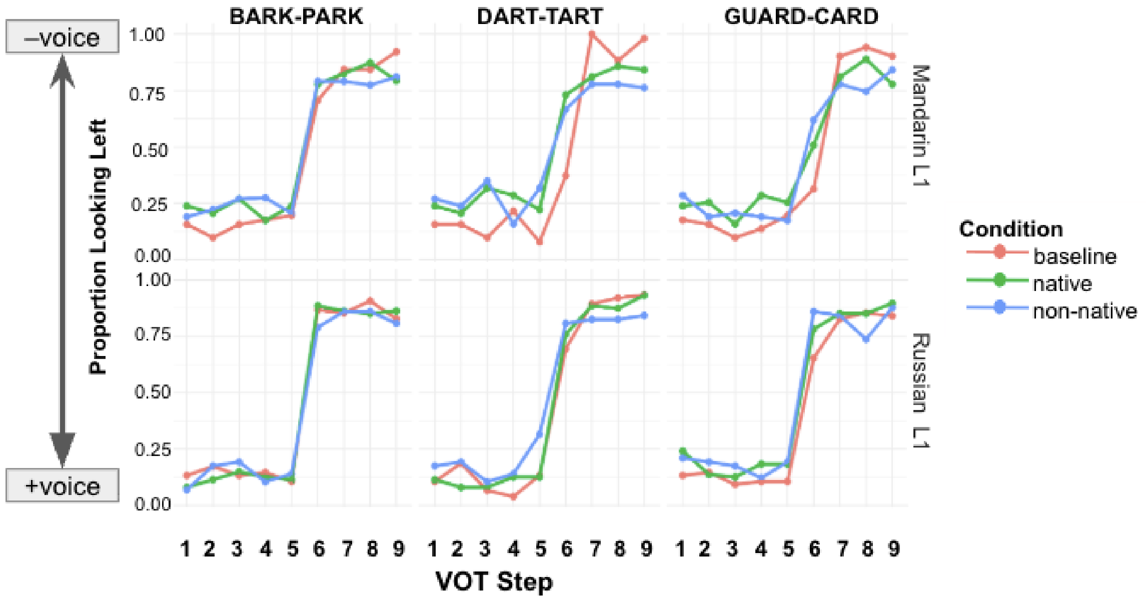


Figure 4.9: Proportion of first looks to the left by VOT step, condition, and listener group. Line plot showing mean voiceless categorization (first leftward gaze) across VOT steps 1–9. Top row: Chinese listeners; bottom row: Russian listeners.

We fit six generalized linear mixed-effects models (GLMMs), one for each listener `group` × `PoA` subset. The model structure was:

```
glmer(first_looked_left  ∼  vot_step * condition + (1 | id),
                            family = binomial)
```

Here, `vot_step` was modeled as a numeric predictor, the `condition` variable included baseline, native, and non-native guises, and a random intercept by `id` accounted for participant variability.

Chinese Listeners

For Chinese listeners, gaze behavior followed a consistent VOT-driven categorical pattern across all three PoAs. In *BARK–PARK*, the shift toward voiceless categorization began as early as Step 6, with a significant increase in looks to the left ($\beta = 2.60$, $p < .001$), followed by steep increases through Steps 7–9. Similarly, *DART–TART* showed a marked rise in voiceless categorization beginning at Step 6 ($\beta = 1.18$, $p = .015$), with peak responses by Step 9 ($\beta = 5.68$, $p < .001$). In *GUARD–CARD*, voiceless categorization emerged strongly from Step 7 onward ($\beta \geq 3.83$, $p < .001$), reflecting robust category boundaries across all contrast sets. The slightly later emergence of voiceless categorization in the velar series (*GUARD–CARD*) likely reflects cross-linguistic and acoustic factors. Velar stops generally exhibit longer and more variable VOTs than labial or alveolar stops (Lisker and Abramson, 1964; Cho and Ladefoged, 1999), which may delay the perceptual boundary. Additionally, compared to Russian listeners, Mandarin listeners' experience with longer VOTs in both Mandarin and English could further reinforce this rightward boundary shift observed for the velar voicing contrast.

Social guise effects emerged primarily at the higher VOT steps. In *BARK–PARK* and *DART–TART*, Step 9 was significantly less likely to elicit a voiceless (left) response when the talker was presented as native (*BARK–PARK*: $\beta = -1.65$, $p = .034$; *DART–TART*: $\beta = -2.79$, $p = .018$), indicating that native guises suppressed voiceless categorization even when the acoustic cue was unambiguous. In *DART–TART* and *GUARD–CARD*, similar suppression effects occurred under the non-native guise, particularly at Step 8 (*DART–TART*: $\beta = -1.47$, $p = .041$; *GUARD–CARD*: $\beta = -2.35$, $p = .004$). See Table 4.2 below for a summary of all significant effects from the three GLMM models, one for each PoA.

Table 4.2: Significant fixed effects for Chinese listeners (RQ2).

Model (PoA)	Effect Term	Estimate	SE	z	*p*-value
CN_ BARK–PARK					
	`vot_step6`	2.60	0.50	5.24	<.001
	`vot_step7`	3.42	0.55	6.23	<.001
	`vot_step8`	3.42	0.55	6.23	<.001
	`vot_step9`	4.21	0.65	6.46	<.001
	`vot_step9` × `native`	–1.65	0.78	–2.12	.034
CN_ DART–TART					
	`vot_step6`	1.18	0.49	2.43	.015
	`vot_step8`	3.77	0.59	6.43	<.001
	`vot_step9`	5.68	1.09	5.23	<.001
	`vot_step9` × `native`	–2.79	1.18	–2.37	.018
	`vot_step8` × `non-native`	–1.47	0.72	–2.05	.041
	`vot_step9` × `non-native`	–3.47	1.16	–2.99	.003
CN_ GUARD–CARD					
	`vot_step7`	3.83	0.60	6.36	<.001
	`vot_step8`	4.39	0.70	6.23	<.001
	`vot_step9`	3.83	0.60	6.36	<.001
	`vot_step9` × `native`	–1.37	0.74	–1.85	.064
	`vot_step7` × `non-native`	–1.62	0.73	–2.21	.027
	`vot_step8` × `non-native`	–2.35	0.81	–2.90	.004

Russian Listeners

For Russian listeners, voiceless categorization began to emerge significantly at VOT Step 6 across all three PoAs. In *BARK–PARK*, Step 6 marked the onset of a significant shift ($\beta = 3.918$, $p < .001$), with similarly strong effects observed at Steps 7–9 ($\beta > 3.5$, all $p < .001$), indicating a sharp, early boundary in categorization. A comparable pattern was found in *DART–TART*, with a significant increase in leftward looks already at Step 6 ($\beta = 3.006$, $p < .001$), followed by a pronounced rise at Steps 7 through 9 ($\beta > 4.3$, all $p < .001$). In *GUARD–CARD*, voiceless categorization also became significant at Step 6 ($\beta = 2.649$, $p < .001$) and remained strong across higher steps.

Although the main effects of `condition` were not significant, two specific interaction terms reached significance (Table 4.3). In *DART–TART*, a significant interaction at Step 9 with the non-native guise ($\beta = -1.549$, $p = .048$) indicated reduced voiceless categorization in this condition, despite the strong acoustic cue. A similar suppression effect appeared in *GUARD–CARD* at Step 8 ($\beta = -1.342$, $p = .043$), also limited to the non-native guise. No significant interactions emerged for the native guise in any PoA. These results suggest that while VOT was the dominant driver of perceptual categorization, non-native guise could attenuate voiceless responses at high VOT steps in some contexts, particularly in mid-to-late phases of categorization in *DART–TART* and *GUARD–CARD*. No such modulation was observed in *BARK–PARK*.

Table 4.3: Significant fixed effects for Russian listeners (RQ2).

Model (PoA)	Effect Term	Estimate	SE	z	p-value
RU_ BARK–PARK					
	`vot_step6`	3.918	0.492	7.971	<.001
	`vot_step7`	3.804	0.482	7.889	<.001
	`vot_step8`	4.329	0.533	8.118	<.001
	`vot_step9`	3.599	0.468	7.695	<.001
RU_ DART–TART					
	(Intercept)	-2.169	0.383	-5.668	<.001
	`vot_step6`	3.006	0.455	6.609	<.001
	`vot_step7`	4.338	0.534	8.121	<.001
	`vot_step8`	4.658	0.572	8.148	<.001
	`vot_step9`	4.856	0.600	8.094	<.001
	`vot_step9` × `non-native`	-1.549	0.785	-1.974	.048
RU_ GUARD–CARD					
	`vot_step6`	2.649	0.430	6.169	<.001
	`vot_step7`	3.630	0.471	7.712	<.001
	`vot_step8`	3.837	0.485	7.905	<.001
	`vot_step9`	3.730	0.477	7.812	<.001
	`vot_step8` × `non-native`	-1.342	0.663	-2.024	.043

Across both listener groups, first look direction revealed consistent VOT-driven categorization, with perceptual boundaries emerging at Step 6 and strengthening sharply through Steps 7–9. However, social guise effects diverged by group and context: Chinese listeners showed consistent suppression of voiceless categorization at high VOT steps under both native and non-native guises in all contrasts, whereas Russian listeners only showed non-native suppression in a limited set of contrasts.

Because these full-model results span the entire continuum, social modulation effects can be diffuse or overlap with strong acoustic cues. To pinpoint where social guise effects are perceptually meaningful, we next isolate analysis at each group's ambiguous VOT step.

Social Effects at Ambiguous Step

To more precisely assess where social guise exerts perceptual influence, I conducted a focused analysis at each group's ambiguous VOT step, defined as the point on the continuum where voiceless categorization was most uncertain (i.e., closest to 50%). These steps were initially identified using model-predicted categorization probabilities from the full GLMMs (see Table 4.4). In most cases, this yielded a perceptually plausible midpoint (e.g., Step 5 or 6). However, in *RU_BARK–PARK*, the model selected Step 9 as the closest to 50% voiceless response, despite this step being acoustically unambiguous and far from the expected perceptual boundary. To maintain interpretive consistency and ensure that true ambiguity can be captured, I manually adjusted the ambiguous step for *RU_BARK–PARK* to Step 6, which exhibited the second-closest predicted categorization rate and better aligned with the acoustic and perceptual midpoint.

Table 4.4: Ambiguous VOT step and predicted voiceless categorization by `group` × `POA`.

`Group` × `POA`	**Ambiguous Step**	**Predicted P (Voiceless)**
`CN_BARK-PARK`	6	0.762
`CN_DART-TART`	5	0.592
`CN_GUARD-CARD`	6	0.480
`RU_BARK-PARK`	6	0.843
`RU_DART-TART`	6	0.758
`RU_GUARD-CARD`	5	0.776

I then re-fit GLMMs for each `group` × `POA` subset, limiting the data to the relevant ambiguous step and testing whether **social guise (condition)** modulated the likelihood of a first leftward (voiceless) gaze. Significant guise effects were observed in three subset models:

- **Chinese listeners** showed enhanced voiceless categorization under both native and non-native guises:
    - In *DART-TART*, compared to baseline, the odds of a leftward gaze increased significantly under both **native** (β = 2.19, p = .004) and **non-native** (β = 1.78, p = .016) conditions.
    - In *GUARD-CARD*, both **native** (β = 0.85, p = .046) and **non-native** (β = 1.32, p = .002) guises similarly boosted voiceless looks relative to baseline.
- **Russian listeners** showed a more selective effect: only in *GUARD-CARD*, the **non-native guise** significantly increased voiceless categorization (β = 1.45, p = .011). No significant native guise effects were found in any PoA for this group.

Table 4.5: Significant guise effects at ambiguous VOT step (by listener group × POA).

Group × POA	**Term**	**Estimate**	**SE**	**z**	***p*-value**
CN_DART-TART	native	2.192	0.754	2.907	.004
CN_DART-TART	non-native	1.777	0.736	2.413	.016
CN_GUARD-CARD	native	0.845	0.423	1.996	.046
CN_GUARD-CARD	non-native	1.316	0.433	3.041	.002
RU_GUARD-CARD	non-native	1.452	0.570	2.544	.011

These patterns are visualized in Figure 4.10, which plots the proportion of first leftward looks (voiceless categorizations) by condition at each group's ambiguous step. Among Chinese listeners in *DART–TART*, only 37%[1] of baseline trials elicited a leftward gaze, compared to 73% in the native guise and 67% in the non-native

[1] See Figure G.1 in Appendix G for descriptive statistics of Voiceless Categorization by Condition for how these percentages were derived.

guise. In *GUARD–CARD*, the voiceless response rate rose from 31% (baseline) to 51% (native) and 62% (non-native). For Russian listeners, baseline voiceless categorization in *GUARD–CARD* was already relatively high (65%) and increased further under social guise (78% in native, 86% in non-native), suggesting social amplification of an existing bias.

These results support the idea that social information can selectively shape perceptual decisions at points of maximal ambiguity, though the direction and magnitude of this effect differ by listener group and lexical context.

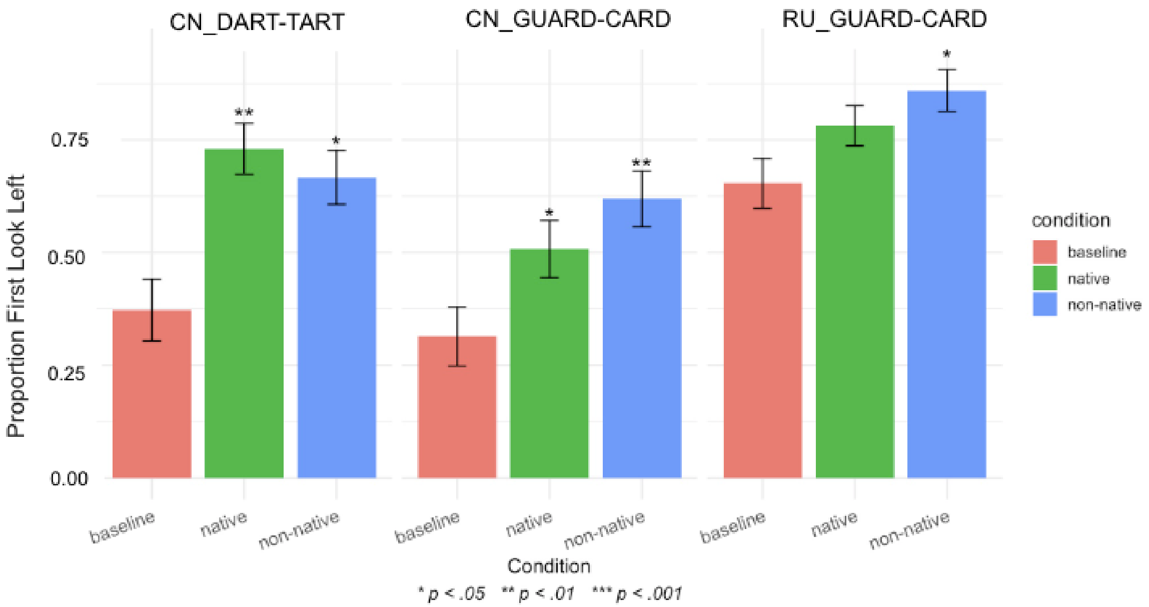


Figure 4.10: Proportion of voiceless categorizations by condition at ambiguous VOT step 6.

4.4.3 Social Modulation on Left Look Onset Latency

Interpreting Latency as a Cue to Early Commitment

Onset latency to the first leftward gaze provides a fine-grained index of how quickly participants commit to a voiceless interpretation. Unlike the binary `first_looked_left` indicator, this measure captures response timing and is sensitive to subtle shifts in real-time categorization. If social guise influences perception, we expect listeners to initiate leftward looks faster or slower depending on the talker's perceived identity. Below we outline the modeling procedure and summarize group-level results across and within VOT steps.

I fit separate linear mixed-effects models for each Listener Group × POA subset using the formula: `lmer(log(first_looks_left_time_ms)` ∼ `condition * vot_step + (1 | id))`. We treat VOT step (as a numeric predictor) and condition (baseline, native, non-native) as fixed effects, with participant as a random intercept.

Across all three continua, both Chinese and Russian listeners showed reliably faster leftward gaze onset as VOT increased. Figure 4.11 plots these patterns by VOT step and condition.

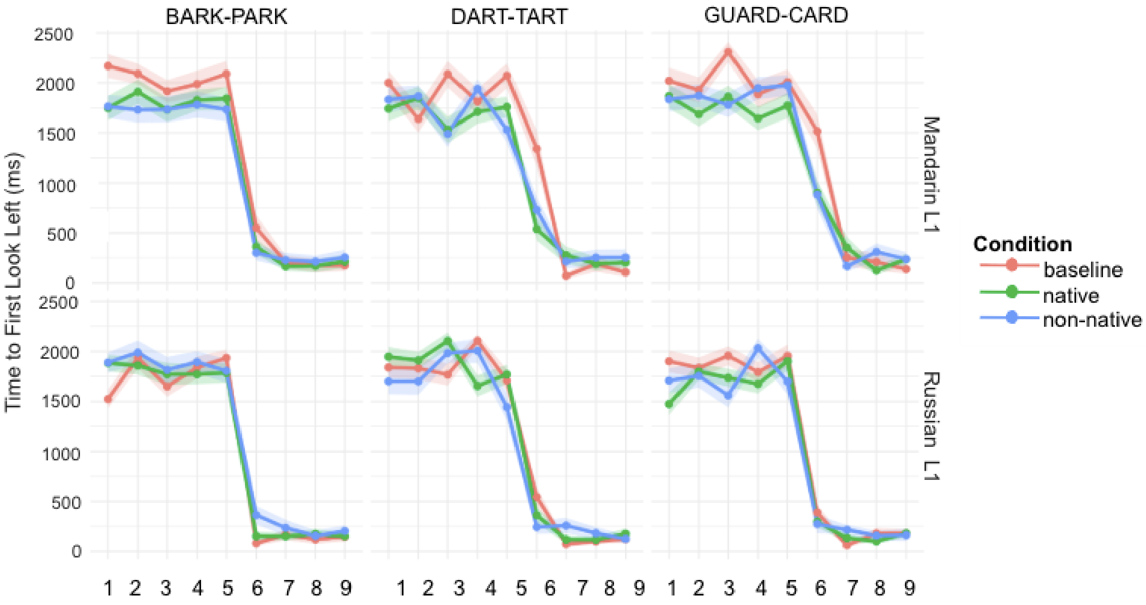


Figure 4.11: Mean onset time to first leftward gaze (ms) across VOT steps. Lines represent condition means; ribbons show ±1 SE. Top row: Chinese listeners; bottom row: Russian listeners.

For Chinese listeners (Table 4.6), each one-step increase in VOT led to a 19% faster look in *BARK-PARK*, 13% faster in *DART-TART*, and 13% faster in *GUARD-CARD* ($p < 0.001$). Additionally, a native guise in *DART-TART* produced a 6.5% further speeding ($p = 0.036$), demonstrating that the presence of an American talker can further accelerate voiceless-cue commitment, but only in selective POA and lexical contexts.

Russian listeners (Table 4.7) exhibited comparable speeding effects across VOT steps: 17% faster per step increase in *BARK-PARK* ($p < 0.001$), 20% in *DART-TART* ($p < 0.001$), and 19% in *GUARD-CARD* ($p < .001$). However, no significant social guise effects were observed in this group.

Table 4.6: Significant effects in left look onset latency of Chinese listeners.

POA	**Term**	**Estimate**	**SE**	***t*-value**	***p*-value**
BARK-PARK	`vot_step`	-0.208	0.0218	-9.50	< .001
DART-TART	`vot_step`	-0.140	0.0252	-5.54	< .001
DART-TART	`native : vot_step`	-0.0676	0.0321	-2.10	.036
GUARD-CARD	`vot_step`	-0.138	0.0202	-6.85	< .001

Table 4.7: Significant effects in left look onset latency of Russian listeners.

POA	**Term**	**Estimate**	**SE**	***t*-value**	***p*-value**
BARK-PARK	`vot_step`	-0.183	0.0180	-10.2	< .001
DART-TART	`vot_step`	-0.219	0.0173	-12.6	< .001
GUARD-CARD	`vot_step`	-0.215	0.0177	-12.1	< .001

Latency Effect at Ambiguous Step

To isolate where social guise most strongly modulates perceptual timing, we focused on VOT Step 6[2], the point of maximum acoustic ambiguity. I fit separate models per

[2] While ambiguous-step analyses in Section 4.4.2 were based on model-estimated voiceless categorization rates (Table 4.5), latency and total-duration measures in Sections 4.4.3

subset using only Step 6 trials. Only the BARK-PARK continuum yielded significant social effects (Table 4.8; Figure 4.12):

- For Chinese listeners, leftward looks under the native guise were initiated 62.5% faster than baseline ($\beta = -0.981$, $SE = 0.448$, $t = -2.19$, $p = 0.040$).
- For Russian listeners, the non-native guise (Chinese) slowed leftward gaze onset by 106% ($\beta = +0.725$, $SE = 0.355$, $t = +2.04$, $p = 0.049$).

Table 4.8: Significant social guise effects on voiceless look onset latency at ambiguous VOT step 6.

Listener × POA	Term	Estimate	SE	t-value	p-value
`CN_BARK-PARK`	native	-0.9810	0.4478	-2.19	.040
`RU_BARK-PARK`	non-native	+0.7253	0.3549	+2.04	.049

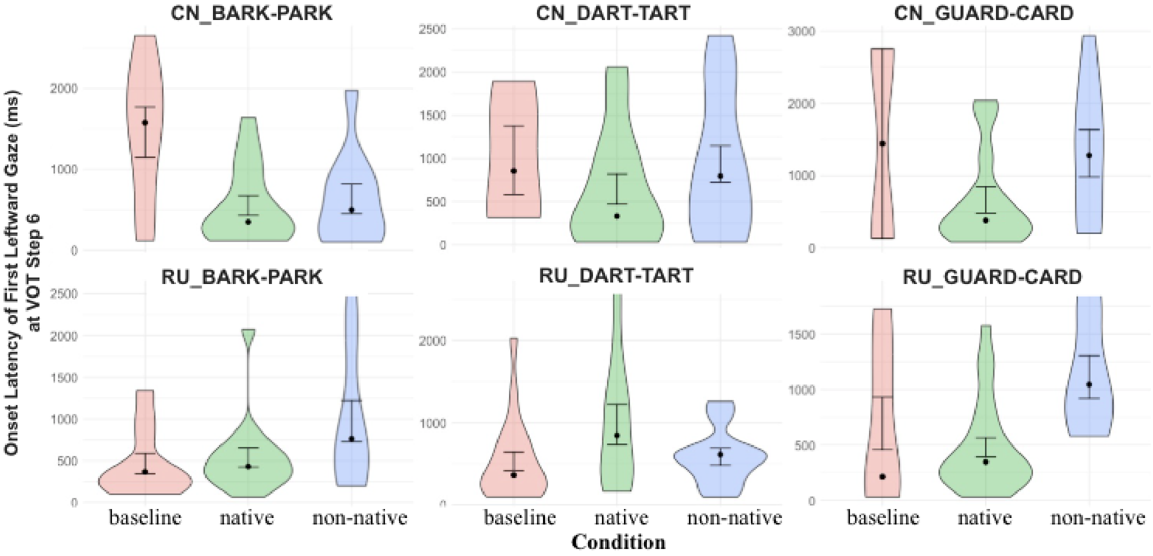


Figure 4.12: Left gaze onset latency at ambiguous VOT step 6 by condition.

and 4.4.4 were analyzed uniformly at VOT Step 6. This choice reflects Step 6's central position in the 9-step continuum and its role as the most acoustically ambiguous point, independent of binary categorization outcomes.

No other subset showed significant effects. These patterns are visualized in Figure 4.12's violins, with a downward shift for the native guise (American) among Chinese listeners (green, top-left) and an upward shift for the non-native guise (Russian) among Russian listeners (blue, bottom-left).

4.4.4 Social Modulation on Total Leftward Duration Time.

Whereas the previous two sections focused on initial gaze decision behavior, such as the probability of the first off-center look being left and the latency to initiate a voiceless-directed gaze, the present analysis considers total duration of gaze to the left as a measure of cumulative perceptual commitment. Specifically, we interpret longer total leftward gaze time as evidence of sustained attention toward the voiceless interpretation across the course of a trial. Unlike earlier metrics that capture initial bias or hesitation, total leftward duration reflects the depth and persistence of voiceless categorization, offering insight into how social guise continues to shape perception over time. An overview of total leftward gaze duration patterns by VOT step, condition, and listener group is shown in Figure 4.13.

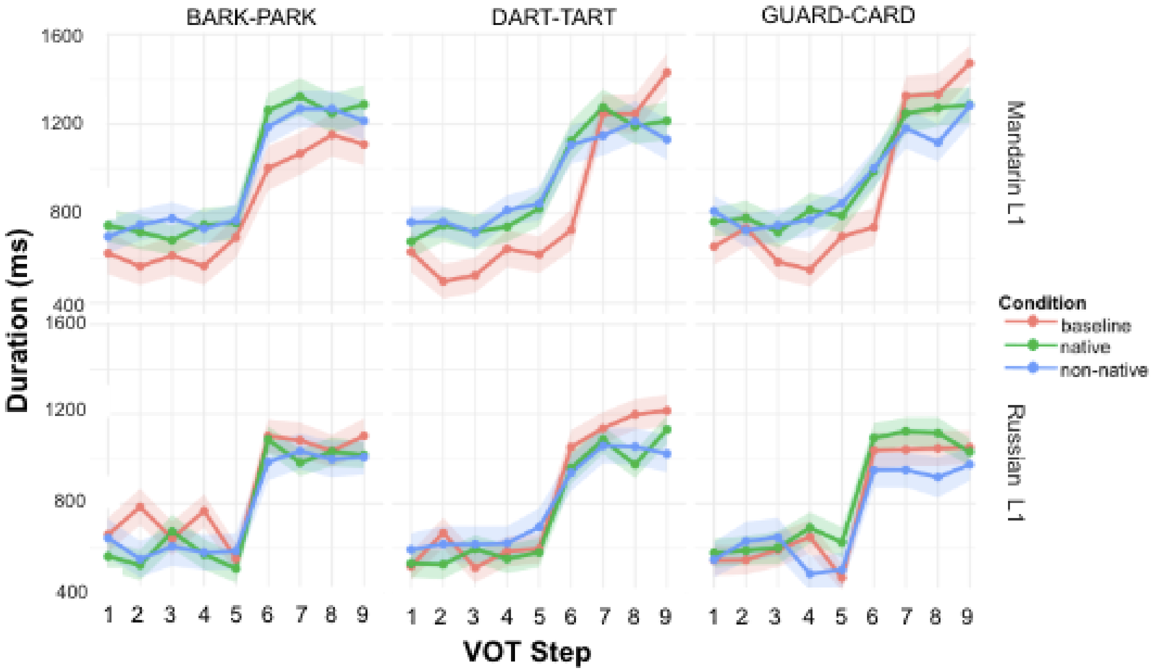


Figure 4.13: Total duration of leftward gaze by VOT step and condition.

Social Effects Across All VOT Steps

To investigate how social guise modulated voicing perception throughout the VOT continuum, we modeled total leftward gaze duration across all nine VOT steps for each listener `group` and `POA` pair. Separate linear mixed-effects models were fit to each of the six subsets, with `vot_step` treated as a factor and condition (`baseline, native, non-native`) as the social manipulation. The dependent variable was the log-transformed total duration of gaze to the left, interpreted as increased voiceless categorization. Each model included random intercepts for participants.

The analysis revealed that social guise influenced voiceless perception in four of the six models (Table 4.9), though the effects were not uniform across the continuum. In some cases, guise exerted a broad influence; in others, its effects were sharply localized to a specific VOT step.

Table 4.9: Significant effects of social guise on leftward gaze duration across all VOT steps.

Listener × POA	Term	Estimate	SE	t-value	p-value
`CN_BARK-PARK`	`native`	0.478	0.241	1.98	.049
`CN_BARK-PARK`	`non-native : vot_step9`	-0.767	0.288	-2.66	.008
`CN_GUARD-CARD`	`non-native : vot_step9`	-0.622	0.302	-2.06	.040
`RU_DART-TART`	`native : vot_step3`	0.740	0.272	2.72	.007

For Chinese listeners, native guises led to a general increase in voiceless categorization in the *BARK-PARK* continuum, reflected in a significant main effect of condition ($\beta = 0.478$, $p = 0.049$). This pattern suggests an overall perceptual bias favoring voiceless interpretations under native social expectations. In contrast, step-specific interaction effects emerged in the *DART-TART* and *GUARD-CARD* continua. Under the non-native guise (Chinese guise), Chinese listeners showed reduced voiceless categorization at Step 9, despite the acoustic signal being highly voiceless. This pattern suggests that strong social cues may override acoustic certainty and bias listeners away from default perceptual mappings (*DART-TART*: $\beta = -0.767$, $p = 0.008$; *GUARD-CARD*: $\beta = -0.622$, $p = 0.040$).

Among Russian listeners, one interaction reached significance in the DART-TART continuum: under the native guise, listeners were more likely to look toward the voiceless target at Step 3, where the VOT value was relatively ambiguous ($\beta = 0.740$, $p = 0.007$). This finding suggests that social cues may sharpen voicing categorization at the early boundary for Russian listeners when expectations align with the guise.

Social Effects At Ambiguous Step 6

Because social guise effects across the full VOT continuum were variable in direction and location, we conducted a focused analysis at Step 6, where the acoustic signal is maximally ambiguous and social expectations are most likely to guide categorization. We fit six separate linear mixed-effects models (one per listener `group` × `POA` subset), predicting the log-transformed total leftward gaze duration as a function of `vot_step` and `condition`, with random intercepts for participants. In this target analysis, only one significant social effect emerged: `CN_GUARD-CARD`. Under the non-native guise, Chinese listeners in the GUARD-CARD continuum prolonged their total gaze on the voiceless target 65% relative to baseline ($\beta = +0.50$, $SE = 0.25$, $t = 2.02$, $p < 0.05$ *), suggesting that social expectations can override acoustic ambiguity to strengthen a perceptual commitment. This effect is visualized in Figure 4.14, which shows the increase in total leftward gaze duration under the non-native condition at Step 6, compared to both baseline and native guises. No other condition or continuum reached significance.

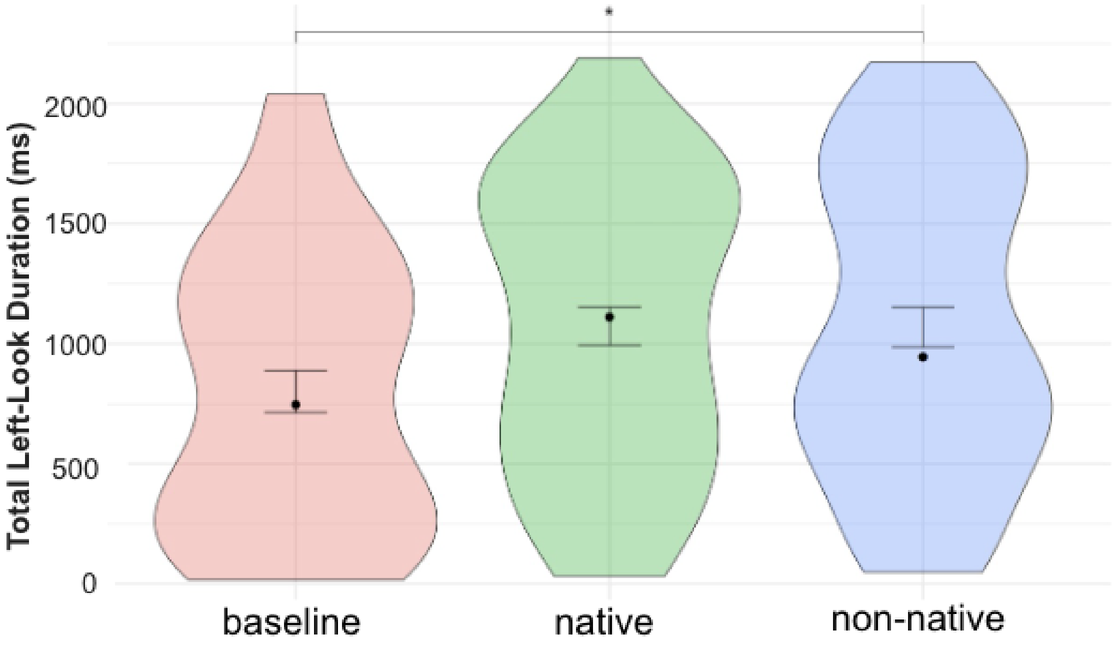


Figure 4.14: Total leftward gaze duration at ambiguous VOT step 6 in `CN_GUARD-CARD` condition.

4.5 Results

4.5.1 RQ1: Feasibility of AEM Paradigm

Baseline trials were used to determine whether the Anticipatory Eye Movement (AEM) paradigm reliably captures categorical perception of voicing. We assessed participants' gaze behavior on unambiguous VOT tokens—Step 1 (voiced) and Step 9 (voiceless)—and calculated accuracy based on whether their first off-center gaze aligned with the correct phonological target. After excluding low-quality recordings, 93 sessions remained technically usable. Among these, 87 sessions met the predefined 70% accuracy threshold, resulting in a 93.55% accuracy-based retention rate from usable sessions and 85.29% from all 102 recordings. This high rate of perceptually reliable behavior across listener groups confirms that the AEM paradigm is an effective tool for tracking real-time phonological categorization.

4.5.2 Q2: Influence of Social Guise on Gaze-Based Categorization

The next question asked whether listeners' gaze behavior, when categorizing voicing contrasts on VOT continua, is modulated by social guise, and if so, how. Across all three gaze metrics (first look direction, onset latency, and total gaze duration), we tested both the presence of an effect (i.e., native or non-native guise vs. baseline) and the type of effect (i.e., native vs. non-native).

Presence of Social Guise Effect

Overall, social guise influenced gaze-based categorization in several contexts, but effects were measure-dependent, POA-specific, and more consistent among Chinese than Russian listeners.

In first look direction, Chinese listeners showed suppression of voiceless categorization at high VOT steps under both native and non-native guises. Russian listeners, by contrast, exhibited fewer effects; only non-native guises reduced voiceless responses, and only in specific continua. These results suggest that the presence of social information can dampen voiceless categorization, particularly among Chinese listeners who may be more sensitive to explicit social cues in unfamiliar speech contexts.

In onset latency, Chinese listeners responded more quickly to voiceless targets as VOT increased. Notably, under the native guise in DART–TART, they exhibited significantly faster gaze onsets, supporting the prediction that social guise can accelerate perceptual commitment. Russian listeners again showed no clear effects of guise on timing, reinforcing the idea that social modulation is more pronounced in the Chinese group.

In total duration, social guise also impacted sustained attention. Chinese listeners exhibited longer leftward gaze durations under the native guise in *BARK–PARK*, and shorter durations under the non-native guise in *DART–TART* and *GUARD–CARD*. These effects were amplified at ambiguous VOT steps, where the acoustic signal provided less clarity. Among Russian listeners, only one condition showed a significant duration effect: increased voiceless gaze time under the native guise in *DART–TART*.

Collectively, these findings support the alternative hypothesis (H_1): the presence of social guise affects online perceptual behavior. While this was not universally true across all groups and conditions, the clearest effects emerged in ambiguous contexts, and were primarily observed in the Chinese group.

Type of Social Guise Effect

We also asked whether the type of guise—native vs. non-native—elicited different perceptual outcomes. Here, the evidence is more selective but still meaningful.

Chinese listeners showed differences between native and non-native guises in both first look direction and duration, particularly in *GUARD–CARD* and *DART–TART*. For example, at ambiguous steps, voiceless responses were significantly more frequent and sustained in the native condition than in the non-native one, suggesting a stronger expectation–cue alignment for native guises. These effects imply that talker identity shapes not only whether social information is used, but how it is interpreted.

Russian listeners, by contrast, showed minimal differentiation between native and non-native guises. When social effects appeared, they were limited to non-native suppression and did not involve divergent responses based on guise type. This asymmetry suggests that Russian listeners may have weaker or less flexible social-perceptual mappings in this context, or that the guises used were less meaningful to them than to the Chinese group.

4.5.3 Do Social Guises Activate L1- or L2-Specific Phonological Expectations?

The findings across gaze measures provide partial support for the idea that a talker's social guise activates language-specific phonological expectations, aligning listeners' perceptual boundaries with the L1 or L2 of the talker.

For Chinese listeners, the results were broadly consistent with predictions. Across all three gaze measures, listeners behaved as if they expected different voicing boundaries depending on the talker's social identity. Under the Chinese guise (non-native), participants were less likely to interpret ambiguous tokens as voiceless, consistent with Mandarin's long-lag VOT system, where longer VOTs are needed to categorize a stop as voiceless. This manifested as:

- **Fewer first leftward looks** in first-look data (especially at high VOT steps in *GUARD–CARD* and *DART–TART*);
- **Slower gaze onsets** to the voiceless target in ambiguous *BARK–PARK* tokens;
- **Shorter total gaze durations** toward voiceless targets, suggesting a weaker commitment to voicelessness.

In contrast, under the American guise (native), Chinese listeners consistently showed greater voiceless categorization—faster gaze onsets, more first leftward looks, and longer total gaze durations—suggesting that they were aligning their perceptual expectations with English phonological norms. These effects were strongest in ambiguous trials (e.g., Step 6), where top-down social expectations have the most room to influence perception.

For Russian listeners, the prediction was that the Russian guise (non-native) would activate a short-lag voicing boundary, making ambiguous tokens more likely to be perceived as voiceless, while the American guise (native) would yield a more conservative, English-like categorization boundary. However, this pattern was not strongly supported across measures.

- In first look direction, social guise effects were minimal and inconsistent.
- In onset latency, the only notable effect was slower voiceless gaze onset under the non-native (Russian) guise, opposite to prediction.
- In total gaze duration, the sole significant effect was an increase in voiceless gaze time under the native (American) guise, again contrary to expectations.

This finding suggests that Russian listeners either did not reliably activate their L1 voicing system in response to the guise, or that the guise cues were less perceptually or socially salient to them than to Chinese listeners. It is also possible that Russian listeners, who generally showed strong categorical behavior across all steps, were less susceptible to social modulation overall.

In short, while Chinese listeners exhibited systematic shifts in voicing categorization aligned with guise-based L1/L2 activation, Russian listeners did not. This asymmetry may reflect group-level differences in L2 dominance, social-indexical sensitivity, or exemplar flexibility, and suggests that the impact of social guise on speech perception is not uniform, but shaped by listener-specific factors and language pair dynamics.

4.6 Discussion

This experiment investigated how social cues influence bilingual lexical processing using an audio-visual matched-guise paradigm with anticipatory eye-tracking. Results revealed that visual information about a talker's identity can modulate phonetic category perception in a lexical task, but this effect was group-specific: Chinese L2-English listeners showed shifts in their perceptual boundaries based on the talker's perceived identity, whereas Russian L2-English listeners did not. Below, I discuss these findings in relation to broader concepts of bilingual lexical processing, real-time gaze behavior, sociophonetic expectation effects, and cross-group sociocultural differences.

4.6.1 Social Modulation of Voicing Contrast

Across all three gaze measures, Chinese listeners exhibited significant changes in categorization behavior based on the perceived identity of the talker. Under the native

(American) guise, they were more likely to categorize ambiguous VOT steps as voiceless—consistent with English phonological boundaries. Under the non-native (Chinese) guise, they showed more conservative categorization, aligning with Mandarin's long-lag VOT boundary. This pattern appeared not only in categorical first looks, but also in faster gaze onset and longer voiceless gaze durations under the American guise.

These effects suggest that visual social cues activated language-specific expectations, shifting perceptual boundaries in the direction of the talker's presumed L1. Such findings are consistent with exemplar-based models of speech perception, which propose that socially-indexed expectations shape how ambiguous input is categorized (Hay et al., 2006; Sumner et al., 2014). They also align with prior matched-guise research showing that listeners adjust their interpretation of speech sounds based on race, accent, or region (Rubin, 1992; Casasanto, 2008; Babel, 2012). The present study extends these effects into real-time lexical decision-making, showing that phonological category activation is not immune to top-down social bias, even when listeners are processing words in isolation.

For Russian listeners, however, the pattern was different. We expected that the Russian guise (non-native) might shift perceptual boundaries toward the shorter VOT distinctions typical of Russian. Instead, social guise had little impact on any gaze measure. While some marginal effects emerged, they were not consistent with predictions and did not show strong evidence of either native or non-native modulation. This suggests that Russian listeners, in this task, were less influenced by social expectations in their categorization behavior.

4.6.2 Interpreting Group Differences

Why might Chinese and Russian bilinguals respond differently to the same social manipulation? We offer several speculative but grounded interpretations, which will be further explored in Chapters 5 and 6. One possible explanation lies in raciolinguistic ideologies (Flores and Rosa, 2015; Rosa and Flores, 2017). Although both groups are L2 speakers of English, their social positioning within U.S. racial and linguistic hierarchies is not equal. Russian immigrants, (majority perceived) as white Europeans, may enjoy greater linguistic security and be perceived as more "legitimate" speakers—even when they are not native. This may reduce both their internal motivation to monitor social cues and their external pressure to adapt their expectations based on a talker's identity. Chinese immigrants, by contrast, may be more aware of being perceived as non-native, regardless of fluency, and more attuned to subtle racialized evaluations of speech (Rubin, 1992; McGowan, 2015; Gnevsheva, 2015). Their increased perceptual flexibility may reflect a kind of sociophonetic vigilance, developed through navigating accent-based marginalization (Lippi-Green, 2012).

Additionally, there may be cultural differences in communication style. East Asian communicative norms emphasize attentiveness to context, social roles, and indirectness (Markus and Kitayama, 1991), whereas Russian communication is typically described as more direct and speaker-focused (Dong, 2009; Leech and Larina, 2014). These tendencies could affect whether listeners are inclined to incorporate talker-related cues into early perception. Such interpretations are speculative, but align with our findings: Chinese listeners appeared more responsive to social context and more willing to shift their perceptual boundary when ambiguity was present. Russian listeners relied more heavily on the acoustic signal itself.

Another contributing factor may be familiarity with talker-specific accent patterns. Chinese listeners may have clearer mental representations of what Chinese-accented English sounds like, making it easier to adjust expectations when seeing a Chinese face. Russian listeners may lack that same perceptual mapping for Russian-accented English in this task context. Chapter 6 will explore how accent familiarity and listener beliefs contribute to category shifting.

4.6.3 Methodological Contributions

Beyond theoretical insight, this study demonstrates the viability of anticipatory eye movements (AEM) task as a method for capturing real-time socially modulated gaze behavior in bilinguals. Our validation analyses show that gaze responses reliably align with phonological categories in baseline trials, and our use of matched guises with controlled audio isolates the role of social expectations from acoustic input. This adds to a growing body of work suggesting that gaze is a powerful tool for uncovering subtle perceptual biases (Clayards et al., 2008; Kong and Edwards, 2011, 2015).

4.6.4 Limitations and Future Directions

We note that the observed social effects were strongest in ambiguous trials and among Chinese listeners, suggesting that social guise is most likely to influence perception when bottom-up cues are weak and when listeners have socialized reasons to question their default expectations. However, social effects were not uniformly distributed across all continua: certain word pairs (e.g., bark–park) yielded stronger effects than others (e.g., guard–card), which may reflect item-level factors such as phonetic salience, lexical frequency, or semantic concreteness. Additionally, ambiguity was defined at the group level using either model-estimated probabilities or fixed step values (e.g., Step 6), which may overlook individual variation in where listeners perceive category boundaries.

In Chapter 5, we will examine whether individual listener traits, such as English language proficiency, cognitive tendencies, social orientation, or language experiences, help predict the strength or direction of these effects. Chapter 6 will build a broader theoretical account of how raciolinguistic ideologies and listener identity interact with bilingual speech perception.

Finally, while our use of OpenFace and PCIbex demonstrates the feasibility of a low-cost, scalable AEM paradigm, the process was not without technical barriers. Despite both platforms being open-source, no existing tools or standardized guidelines were available for transforming OpenFace gaze output into interpretable, trial-aligned data for speech perception research. As such, the entire analysis pipeline had to be developed from scratch, in consultation with a computer vision specialist. The workflow was also machine-intensive—processing even a single well-formatted 20-minute video session can itself take more than 20 minutes. Although the processing presented logistical challenges, it also represents a core methodological contribution of the project: the trial-and-error required to build this system positions us to offer a reusable framework for future researchers working with web-based eye-tracking in sociophonetics.

Chapter 5

DISENTANGLING LISTENER DIFFERENCES: TASK DEMAND, ENGLISH ORAL PROFICIENCY, AND AUTISM SPECTRUM TRAITS IN BILINGUAL SPEECH PROCESSING

5.1 Introduction

Speech perception in bilingual populations is marked by considerable complexity and variability, setting bilingual listeners apart from their monolingual counterparts. Unlike monolinguals, bilinguals must constantly navigate the phonetic, lexical, and social demands of multiple language systems, resulting in greater flexibility, but also greater heterogeneity, in their perceptual strategies.

Previous chapters of this dissertation have begun to disentangle how bilinguals process voicing contrasts in English, revealing both group-level patterns and systematic differences between Russian-English and Mandarin-English bilinguals. In the VAS task (Chapter 3), both groups exhibited significant social-cue modulation of voicing perception, but the pattern of these effects diverged: Mandarin-English listeners showed pronounced guise effects, especially in voiced zones and in response to the Russian guise, while Russian-English listeners displayed more complex, bidirectional shifts across acoustic continua, with the strongest modulation prompted by the Mandarin guise. This group divergence was also evident in the AEM task (Chapter 4), where Chinese listeners consistently shifted their perceptual boundaries in line with talker identity, whereas Russian listeners showed minimal and inconsistent social-cue effects across all gaze measures.

Yet, these group-level comparisons and task differences raise important questions: To what extent do such differences truly reflect stable characteristics of listeners' native language groups? Are all Russian bilinguals alike, or all Chinese bilinguals alike, in how they process speech and integrate social cues? Or do individual factors, such as language proficiency and cognitive style, drive meaningful variability within groups that may rival or even outweigh between-group effects?

5.1.1 Motivation

Caution is warranted in drawing broad conclusions about group-level behavior, especially in bilingual populations marked by substantial heterogeneity. Attributing perceptual patterns solely to ethnicity, L1 background, or country of origin risks reinforcing essentialist views and masking the rich variability present within speaker communities. Indeed, contemporary research in sociolinguistics and speech science increasingly highlights the importance of individual differences in language processing. Third wave sociolinguistics, for example, emphasizes the agency of individuals—their positionality, stance, and engagement with sociolinguistic meaning—which is dynamic, fluid, and context-dependent (Eckert, 2018). Treating individuals as agents rather than static samples of fixed demographic categories enables a more nuanced understanding of linguistic behavior and prevents the reinforcement of harmful stereotypes.

On the cognitive side, prior studies have demonstrated that individual variation in cognitive style, language experience, and proficiency can strongly influence speech perception among monolingual and bilingual groups alike. Recent research has shown that abilities such as attentional switching, working memory capacity, and inhibitory control contribute to individual differences in how listeners perceive and produce speech

(Ou et al., 2015; Ou and Law, 2017; Lev-Ari and Peperkamp, 2013). For example, in Cantonese tone merger, stronger attentional control and working memory predict faster, more accurate discrimination and higher-quality perceptual representations (Ou et al., 2015; Ou and Law, 2017). Cognitive processing style, such autism-aligning traits, have also been linked to how listeners normalize for phonotactic variation and respond to context-dependent phonological effects (Yu, 2010; Yu et al., 2011). Yu (2010) found that women with low Autism Spectrum Quotient (AQ) scores normalized less for phonetic coarticulation, illustrating how subtle cognitive differences can shape perceptual strategies.

In bilingual populations, recent work has shown that general cognitive skills also shape cross-linguistic interactions in perception and production. Lev-Ari and Peperkamp (2013) demonstrated that bilinguals with lower inhibitory control exhibited greater cross-language influence: late English–French bilinguals with poorer inhibition produced and perceived English stops in a more French-like manner, indicating greater co-activation of the non-target language. Similarly, Roberts (2012) reviewed evidence that working memory and L2 proficiency impact real-time sentence processing in L2 learners, but found these effects tend to emerge primarily under experimental conditions requiring explicit metalinguistic attention. Under more naturalistic processing conditions, individual variability may play a smaller role.

While these studies collectively demonstrate that cognitive abilities can systematically affect speech and language processing, in the context of the present study, it remains an open question how cognitive and experiential factors shape bilingual listeners' responses when they must navigate the simultaneous demands of top-down social expectations–imposed by our guise manipulations – and bottom-up acoustic processing, where the same yet nuanced VOT continua are repeated under different conditions.

To address these questions, this chapter takes an explicitly individual differences approach. In addition to the group-level comparisons reported in earlier chapters, I administer a battery of objective assessments to each participant, including an oral English proficiency test (Elicited Imitation Task, EIT) and the Autism Spectrum Quotient (AQ). These measures allow me to directly model how language proficiency and cognitive style predict listeners' perceptual boundaries and their flexibility in adapting to social cues — both within and across L1 groups.

Research Questions and Predictions

This chapter addresses the following exploratory questions, each accompanied by specific predictions grounded in prior literature.

1. **Task Comparison.** How do listeners' perceptual boundary shifts in response to guises compare across the AEM and VAS tasks? Are task-driven effects consistent across groups?

   Predictions. Social guise effects will be more robustly captured in the AEM task than in the VAS task. Because the AEM task requires rapid, real-time responses with minimal opportunity for conscious filtering or social desirability bias (Fisher, 1993), listeners are expected to show stronger or more genuine shifts in perceptual boundaries compared to the more reflective, self-monitored responses in the VAS task. Previous research has shown that explicit rating tasks are vulnerable to social desirability and participant self-monitoring (Niedzielski, 1999; Vaughn and Walker, 2024), whereas implicit measures such as eye-tracking more directly reflect spontaneous perceptual adjustments (see (Hay and Drager, 2010)).

2. **Role of English Proficiency.** Does L2 oral proficiency (EIT) account for group-level differences in voicing perception, or do such differences persist even after controlling for proficiency? How does proficiency affect responses within each task?

   Predictions. Listeners with higher EIT scores are anticipated to show greater alignment with English-like voicing boundaries, particularly under the native (American) guise. This prediction is supported by previous work that demonstrates that higher proficiency corresponds to more native-like categorical boundaries at the speech level (Flege, 1995; Flege et al., 1995), and proficiency facilitates flexible adaptation to L2 contexts. Conversely, lower proficiency may result in reduced sensitivity to social cues and less alignment with L2 norms, as perceptual categories may be less firmly established (Roberts, 2012; Lev-Ari and Peperkamp, 2013).

3. **Cognitive Style Effects.** Do socio-cognitive styles as reflected in AQ total and subscales predict listeners' sensitivity to social guise and their flexibility in shifting perceptual boundaries? Are certain cognitive profiles associated with greater or lesser adaptation to social information?

   Predictions. Individuals with lower AQ scores (i.e., less autism-aligned traits) are expected to exhibit greater shifts in response to social guise, consistent with heightened social awareness and a greater reliance on top-down processing (Yu, 2010; Yu et al., 2011). Higher AQ scores, on the other hand, are predicted to correspond to reduced sensitivity to social cues and increased reliance on bottom-up acoustic information, aligning with research linking autistic traits to decreased perceptual flexibility and diminished influence of contextual effect (Stewart and Ota, 2008; Yu, 2010).

Unlike earlier chapters, I frame the above possible outcomes as predictions rather than strict hypotheses to reflect the exploratory nature of this individual-differences approach. Rather than presupposing fixed group effects, predictions acknowledge the continuous, multidimensional variability in our data, where listener group, task, VOT continua, experimental conditions, and individual factors (proficiency, AQ) interact and avoid presupposing directional outcomes. This approach mitigates the ecological fallacy of attributing group-level patterns to all individuals (Robinson, 2009) while allowing emergent patterns to guide interpretation.

5.1.2 Implications and Chapter Roadmap

By integrating both proficiency and cognitive style metrics, this chapter moves beyond essentialist group comparisons, offering a more dynamic account of bilingual speech perception. In the sections that follow, I first present a direct comparison of AEM and VAS performance (Section 5.2), then examine how English proficiency (EIT) relates to categorization patterns in each task (Section 5.3), and finally assess how AQ and its subscales correlate with individual shifts in response to social cues during the AEM task (Section 5.4). Our analytical strategy prioritizes data-driven discovery: in Sections 5.3 and 5.4, I first model how proficiency and cognition shape responses within and across listener groups before examining their interplay with social cues.

5.2 Comparison of Anticipatory Eye Movement and Visual Analogue Scale Performance

5.2.1 Data Preparation and Alignment of AEM and VAS Tasks

To compare the performance of listeners across the Anticipatory Eye Movement (AEM) and Visual Analog Scale (VAS) tasks, I created a merged data frame including only trials and conditions that were directly comparable between tasks.

Variable Selection and Filtering

For the AEM task, I focused exclusively on the gaze metric `first_looked_left`, which reflects the earliest and most anticipatory categorization decision available in the eye-tracking paradigm. This measure was chosen for its conceptual similarity to the overt categorization response provided in the VAS task.

For the VAS data, only responses from the baseline and the first social block were included, consistent with the approach in Chapter 3. Responses to the "unfamiliar non-native guise" (i.e., the non-matching social condition present in VAS but not AEM) are excluded so that both tasks would only include comparable experimental contexts.

Data Merging and Creation of the Combined Data Frame

The AEM and VAS datasets are merged and filtered into a single data frame to facilitate parallel analysis. In this merged dataset, several columns were standardized or newly created to ensure clear comparison:

- `response`: Identifies percentage of voiceless categorization in the VAS task and the percentage of first off-centered look being leftward in the AEM task.

- `exp_condition`: Identifies both the task type and the experimental condition for each trial (e.g., `AEM_baseline`, `VAS_baseline`). This variable has 7 levels, with 3 from AEM and 4 from VAS.

- `stimuli_type`: Indicates whether the stimulus was a lexical item (AEM) or a syllable (VAS).

- `poa_type`: Categorizes place of articulation into three types: bilabial, alveolar, and velar, harmonized across both tasks.

- `task`: Binary indicator for AEM vs. VAS.

- `aligned_guise`*:* A crucial derived variable that codes aligned experimental contrasts between the two paradigms, allowing for direct comparison. For example, in Chinese listeners, `VAS-baseline` is paired with `AEM-baseline` (`aligned_guise = "CN_combined_baseline"`), while `VAS-Mandarin` is paired with `AEM-non-native` (`aligned_guise = "CN_combined_Mandarin"`), and so forth. This alignment ensures that statistical comparisons are made only between theoretically and contextually matched conditions.

Rationale for Guise Re-Alignment

Aligning guises is essential because the two paradigms differed in their experimental manipulations: the VAS task included some conditions that were not present in AEM and vice versa. Creating the aligned_guise variable ensured that comparisons between tasks are restricted to those experimental conditions that are conceptually and operationally parallel, eliminating confounds due to unpaired experimental manipulations.

To visualize overall distributions prior to modeling, Figure 5.1 shows the distribution of voiceless categorizations across the seven experimental conditions, combining both AEM and VAS tasks. The violin shapes reveal task-specific differences in variability, with AEM conditions generally exhibiting more polarized voiceless ratings compared to VAS conditions. As we can see, across both listener groups and the seven guises, the AEM task in general evoked a great range of response distributions. Figure 5.2 complements the violin plot by breaking down responses by VOT step, illustrating that across all steps of the continua, responses to syllables (VAS task) is more categorical than responses to lexical items in the AEM task.

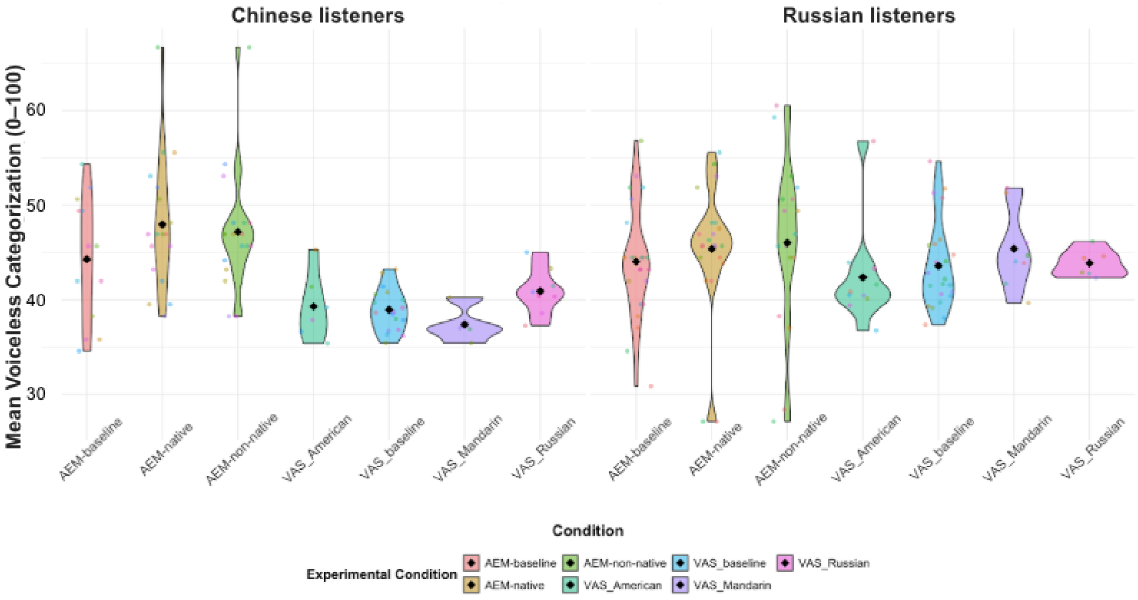


Figure 5.1: Individual and group mean for voiceless ratings across experimental conditions.

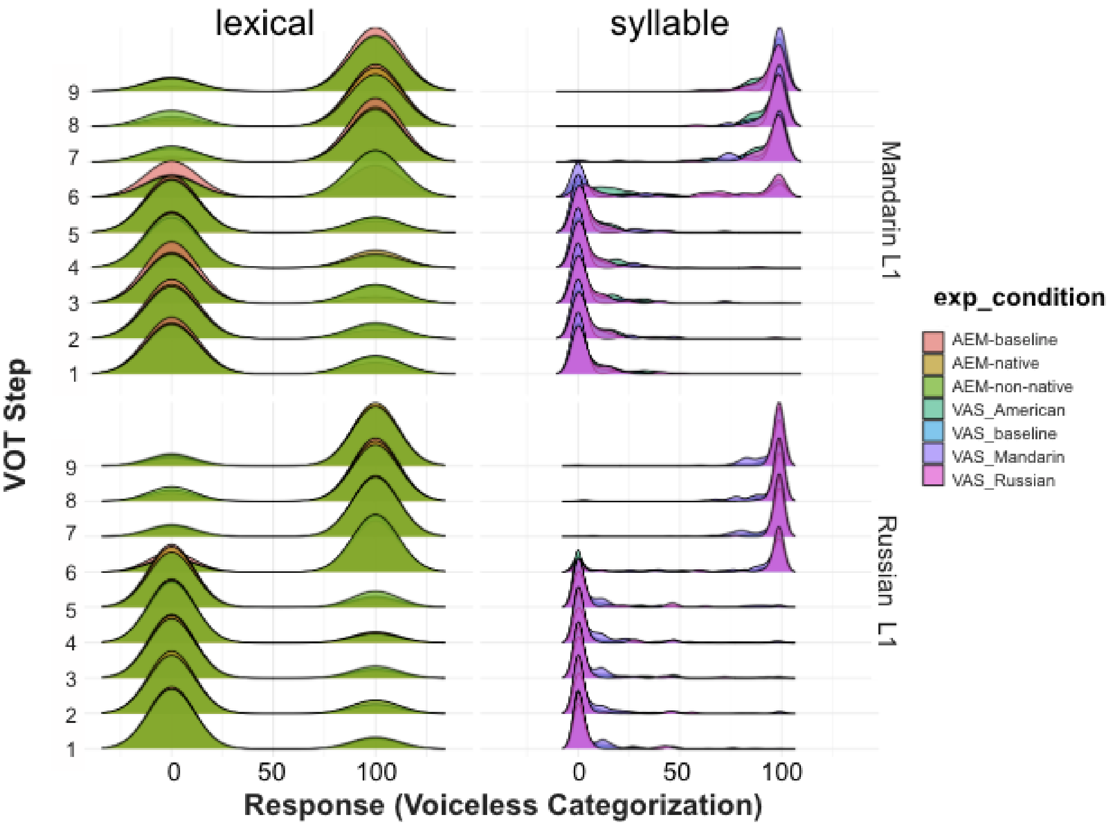


Figure 5.2: Distribution of voicing responses by VOT step and experimental condition.

5.2.2 Statistical Modeling Approach

A subset for each listener group (*Chinese* and *Russian*) is extracted from the combined data frame and filtered to include only trials with a valid `aligned_guise` value. Then, linear mixed-effects models are fit to each subset to predict the voiceless categorization response as a function of VOT step, task (AEM vs. VAS), and their interaction, with random intercepts for participant and place of articulation:

```
lmer(response  ∼  aligned_guise * vot_step * task +
                  (1 | ID_combined) + (1 | poa_type))
```

5.2.3 Results

Chinese Listener Model Results

For Chinese listeners, the model revealed several statistically significant effects of both acoustic and experimental factors on voiceless categorization. As expected, the likelihood of categorizing a stimulus as voiceless increased strongly with each unit increase in VOT step ($p < .001$, $\beta = 14.17$), reflecting reliable sensitivity to this primary acoustic cue. Task type was also a significant predictor: participants were overall more likely to make voiceless (leftward) categorizations during the AEM task than during the VAS task ($p < .001$, $\beta = 14.64$). However, this task effect may be partially attributable to our response metric, `first_looked_left`. Since prior eye-tracking research demonstrates a general bias for people to look left (potentially stemming from habitual scanning patterns in left-to-right readers), this could have inflated the likelihood of voiceless categorizations in the AEM task. In this sense, the observed effect not only reflects a known attentional bias but also lends further validation to our study design, as it aligns with broader literature on task and gaze-based response tendencies.

Beyond these main effects, several significant interactions were observed. There was a robust interaction between task and guise, such that the AEM-VAS difference in voiceless categorization was even larger in the Mandarin guise ($p < .001$, $\beta = 24.99$) and the American guise ($p = .008$, $\beta = 14.23$), relative to the baseline guise. Furthermore, the difference in slope of voiceless categorization across VOT steps (i.e., VOT sensitivity) between AEM and VAS tasks was significantly reduced in Mandarin ($p < .001$, $\beta = -4.13$) and American guises ($p = .012$, $\beta = -2.37$), compared to baseline. This suggests that social information about the talker modulated not only the overall response level but also the degree to which VOT was used to guide categorization across tasks.

Table 5.1: Significant task, VOT, and guise effects in Chinese listeners.

Term	Estimate	SE	t	p-value
(Intercept)	-31.91	2.72	-11.73	$< .001$
vot_step	14.17	0.38	37.73	$< .001$
taskAEM	14.64	3.20	4.57	$< .001$
vot_step : taskAEM	-1.81	0.57	-3.19	.001
taskAEM : aligned_guiseCN_combined_American	14.23	5.34	2.66	.008
taskAEM : aligned_guiseCN_combined_Mand	24.99	5.94	4.21	$< .001$
vot_step : taskAEM : aligned_guiseCN_combined_American	-2.37	0.94	-2.51	$< .001$
vot_step : taskAEM : aligned_guiseCN_combined_Mand	-4.13	1.06	-3.94	$< .001$

The random effects structure captured both between-participant and item-level variability. Specifically, moderate variation across participants ($SD = 2.72$) and place of articulation ($SD = 2.75$) are observed, with the majority of residual variability remaining at the trial level ($SD = 37.0$). This pattern of results is consistent with robust individual differences and some POA effects, but the primary variance in responses is driven by stimulus-level and within-subject factors.

Russian Listener Model Results

For Russian listeners, the model identified several robust effects and interactions between VOT step, task type, and experimental guise (see Table 5.2for all significant fixed effects).

As anticipated, listeners were far more likely to categorize a stimulus as voiceless with increasing VOT step ($p < .001$, $\beta = 14.75$), indicating strong cue sensitivity in both AEM and VAS tasks. There was also a significant main effect of task: Russian participants showed higher overall rates of voiceless responses during the AEM task compared to the VAS task ($p < .001$, $\beta = 15.07$). The main effects of experimental guise were not statistically significant; both the American and Russian-aligned guises did not differ significantly from baseline in overall voiceless categorization rates.

Several interactions, however, point to important task- and guise-related modulation of VOT effects. The negative and significant interaction between VOT step and task ($p < .001$, $\beta = -2.89$) suggests that the effect of increasing VOT on voiceless responses was stronger in the VAS task compared to AEM, mirroring a pattern seen for Chinese listeners. Of particular interest, the interaction between task and the Russian-aligned guise was significant ($p = .029$, $\beta = 10.42$), indicating that in the Russian guise condition, the difference between AEM and VAS tasks was especially pronounced. There was also a marginal three-way interaction between VOT step, task, and Russian-aligned guise ($p = .052$, $\beta = -1.62$), suggesting subtle modulation of cue use by both task and social context, although this effect did not reach conventional significance.

Table 5.2: Significant task, VOT, and guise effects in Russian listeners.

Term	Estimate	Std. Error	t value	p-value
(Intercept)	-30.17	2.28	-13.22	< .001
vot_step	14.75	0.31	47.94	< .001
taskAEM	15.07	2.48	6.09	< .001
vot_step : taskAEM	-2.89	0.44	-6.57	< .001
taskAEM : aligned_guiseRU_combined_Russian	10.42	4.77	2.19	.029
vot_step : taskAEM : aligned_guiseRU_combined_Russ	-1.62	0.84	-1.94	.052

Random effects were substantial at both the participant and place-of-articulation levels. The standard deviation for participant intercepts was 4.05, indicating meaningful between-listener variability in overall response tendencies. Variability across place of articulation ($SD = 2.14$) also suggests that different POAs contributed to response differences, while the residual error ($SD = 35.0$) was similar to the Chinese model, indicating comparable within-group noise across listener groups.

Russian listeners demonstrated strong and consistent use of VOT in voicing categorization, as well as robust AEM task effects and selective interactions with guise, especially in the Russian-aligned condition. These results further reinforce the influence of both perceptual task and social context on speech sound categorization across bilingual listener groups.

In summary, the AEM paradigm elicited stronger shifts under different guises than the VAS ratings did, supporting the prediction that implicit eye movement measures are more sensitive to social cues than explicit judgments. This addresses RQ1 indicating that social guise effects were indeed more robustly captured in AEM, though task differences were nuanced by listener group and guise type. Having seen that task format influences the strength of social effects, I next examine whether listeners' English proficiency can account for some of the differences between individuals and between our two language groups.

5.3 Elicited Imitation Task

5.3.1 Overview, Purpose, and Rationale

The Elicited Imitation Task (EIT) is a short-cut oral proficiency test initially developed in psycholinguistics for assessing oral language proficiency efficiently and objectively. Due to its brevity, reliability, and ease of administration, it has been widely adopted for linguistic research on second language (L2) proficiency. Advantages of the

EIT include its controlled stimuli, objective scoring criteria, and ability to reliably distinguish levels of language proficiency in research settings.

The EIT utilized in this study consists of 30 sentences presented in a fixed order, incrementally increasing in length from seven to nineteen syllables. Participants listened to recordings produced by a female native speaker of American English obtained from the IRIS database (IRIS Database, n.d.) After hearing each sentence, participants received approximately 2.5 seconds of silence, followed by a beep signaling them to repeat the sentence aloud. Each participant had 1.5 times the original sentence duration to complete their response before the next sentence automatically played.

The EIT was conducted primarily in the Linguistics Lab, with some sessions occurring off-campus. It was always administered during the first experimental session, immediately following the VAS task. Participants completed the task alone, without the researcher present, to maximize comfort and minimize effects from the observer's paradox (Labov, 1972). The task lasted approximately 7 minutes and 30 seconds. Recordings were collected primarily using an H4N Zoom recorder, supplemented occasionally by an iPhone app and Zoom to ensure reliable data backup; all recordings were stored in an uncompressed WAV format.

To ensure validity of scoring, two trained undergraduate research assistants, both native speakers of American English, independently scored each participant's recordings using a 4.0 grading scale, referencing rubrics from established sources (Bowden, 2016; Wu and Ortega, 2013; Tracy-Ventura et al., 2014). Any scoring discrepancies were subsequently resolved in collaboration with the Principal Investigator (PI).

Integrating EIT into this study allows examination of whether individual differences in L2 proficiency—independent of native language (L1) background—shape bilingual listeners' perceptual categorization of voicing contrasts on both the Anticipatory Eye Movement (AEM) and Visual Analogue Scale (VAS) tasks. Previous analyses indicated systematic categorization differences between Russian and Chinese bilingual listeners, raising the question of whether these differences are primarily attributable to L1 phonological systems or significantly influenced by variations in English proficiency within our participant sample. Thus, this section explicitly tests whether proficiency, as indexed by EIT scores, explains variance beyond what can be attributed to listeners' native language group membership.

5.3.2 Statistical Approach

To determine the influence of English proficiency (EIT scores) on listeners' perceptual voicing categorization, linear mixed-effects models were implemented. These models enable simultaneous examination of the effects of VOT step, EIT proficiency scores, social guise conditions (Baseline, American, Mandarin, Russian), and listener groups (Chinese and Russian bilinguals) on participants' perceptual responses.

Analyses were structured in two phases. First, four separate models were run for the Anticipatory Eye Movement (AEM) task and the Visual Analogue Scale (VAS) task, examining each listener group independently, employing the following structure:

```
lmer(response  ∼  vot_step * EIT * guise +
                  (1 | ID_combined) + (1 | poa_type))
```

This initial exploration allowed us to evaluate how proficiency affected perceptual categorization within each group and task independently.

Next, two cross-group analyses for each task were conducted by integrating both listener groups within unified models. These comprehensive models included all two-, three-, and four-way interactions involving EIT, listener group, VOT steps, and guise conditions, structured as follows:

```
lmer(response  ~  vot_step * EIT * listener_group * guise +
                  (1 | ID_combined) + (1 | poa_type))
```

The cross-group modeling aimed to explicitly assess whether group differences in perceptual categorization could be explained by differences in English proficiency, or if they persisted even after controlling for proficiency effects.

All statistical procedures were performed using the lme4 package in R, with significance testing conducted via Satterthwaite's method to approximate degrees of freedom. Results are reported below, with detailed statistics summarized in appendices for transparency and reproducibility.

EIT and Real-Time Speech Perception: Proficiency Effects in the AEM Task

Before examining model outcomes, it is useful to first describe the distribution of English oral proficiency scores across the two listener groups and how proficiency levels may relate to voiceless categorization in the AEM task. This analysis includes 40 participants in total: 18 Chinese listeners and 22 Russian listeners. Each participant's EIT score was computed based on their performance on a 30-sentence elicited imitation task (see Section 5.3.1).

The histograms in Figure 5.3 illustrate the distribution of EIT scores for Chinese and Russian listeners separately. Chinese listeners show a moderately wide spread of EIT scores, ranging from approximately 75 to 115, with no single dominant peak. Russian listeners display a comparatively right-skewed distribution, with a larger cluster of scores in the higher proficiency range $(100 - 110)$, suggesting generally stronger English proficiency across this group compared to the Chinese group.

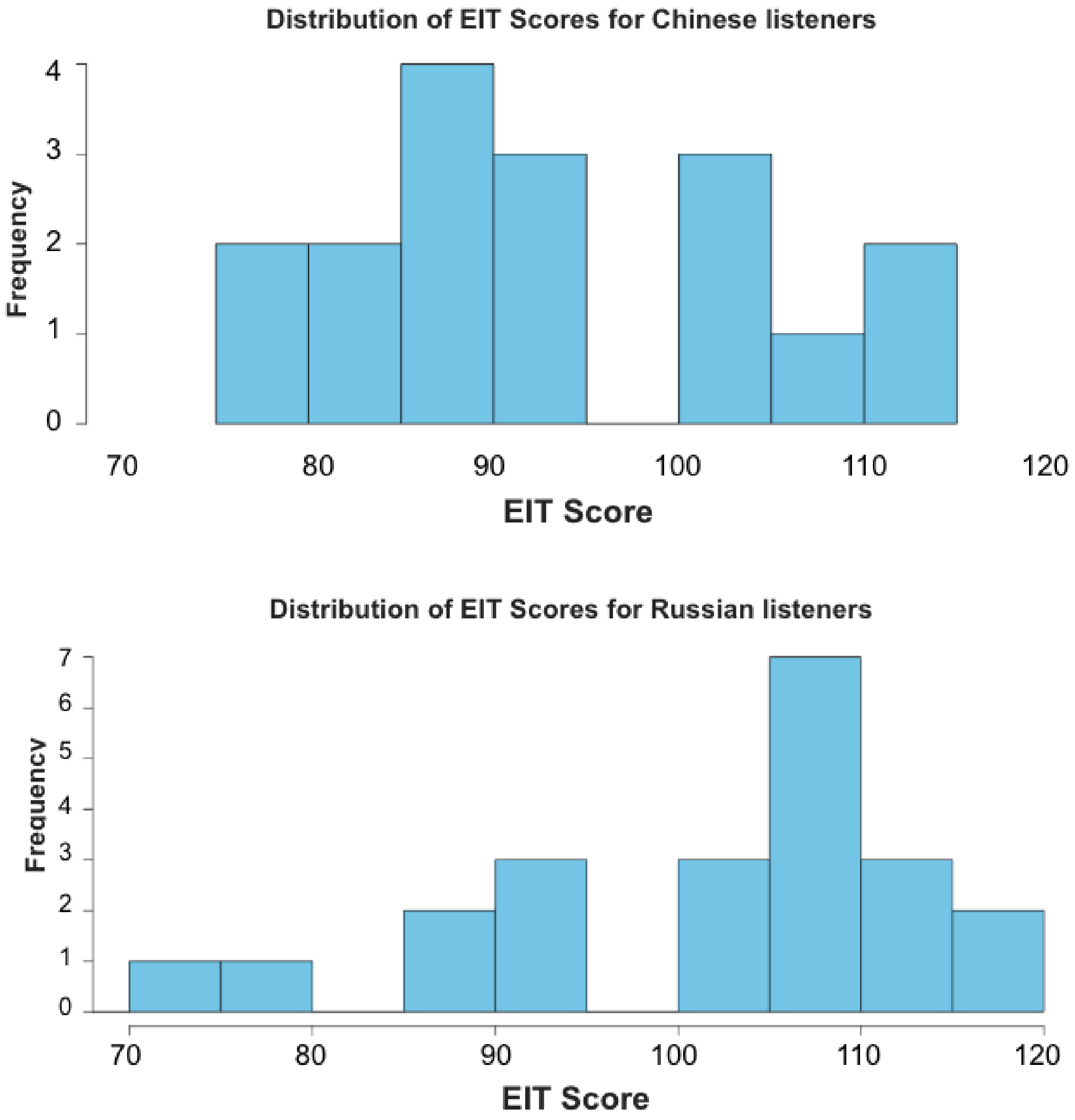


Figure 5.3: Distribution of EIT scores across listener groups. Top: Chinese listeners; bottom: Russian listeners. Note: Bin width = 5 points.

To visualize how proficiency affects perceptual behavior, EIT scores were divided into tertiles based on the full dataset distribution ($N = 40$). The cutoff points used were:

- **Low proficiency**: EIT ≤ 91

- **Mid proficiency**: $91 < \text{EIT} \leq 107$
- **High proficiency**: $\text{EIT} > 107$

This results in three equally sized EIT groups used for visualization. These groupings are plotted in Figure 5.4, which displays mean voiceless categorization across the nine-step VOT continuum (1-9), split by guise condition and listener group.

For Chinese listeners, the baseline condition reveals the clearest pattern: higher proficiency listeners (blue line) show elevated voiceless responses at ambiguous or voiced-like steps (Steps 2-6), suggesting a shift towards English-like perceptual reliance on VOT contrast despite minimal acoustic evidence. In contrast, the American and Mandarin guises show more overlapping trends across EIT levels.

Among Russian listeners, EIT-related differences are most visible in the baseline and American guises. In these conditions, high-proficiency listeners again demonstrate more categorical behavior, with sharper perceptual shifts across the VOT continuum. The blue line rises more steeply, suggesting that greater proficiency may be associated with more decisive (leftward look) voiceless identifications. However, unlike the Chinese group, these differences appear more gradual and less condition-dependent.

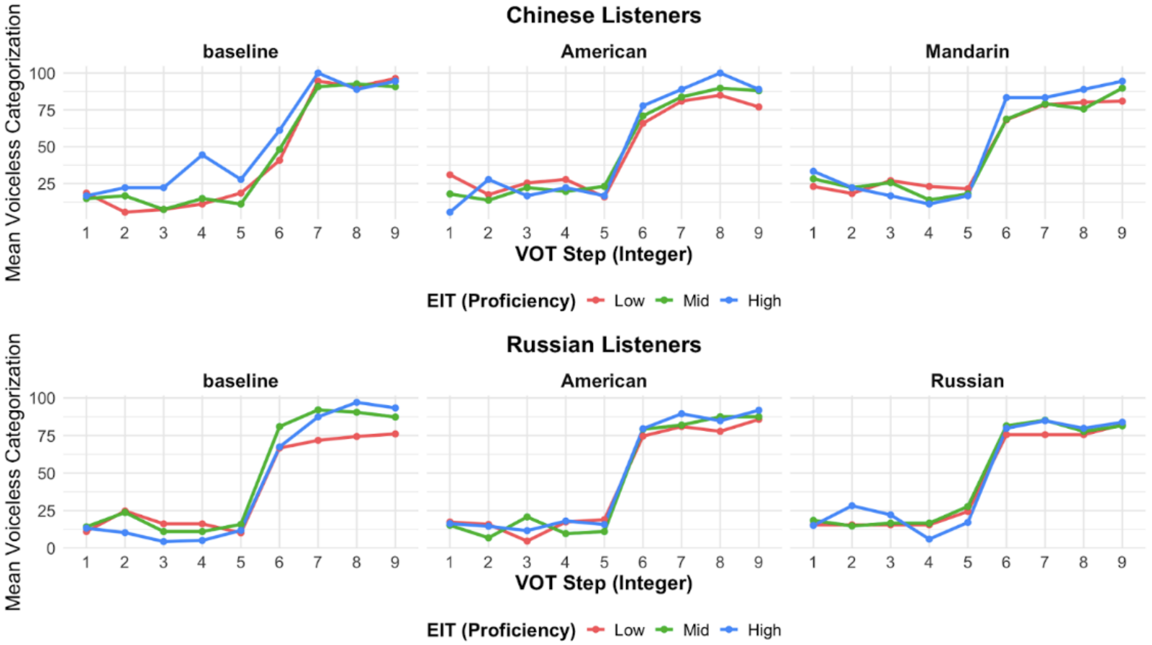


Figure 5.4: AEM categorization by VOT step and English proficiency (EIT). Top: Chinese listeners by guise; bottom: Russian listeners by guise.

Chinese Listeners. The linear mixed-effects analysis for Chinese listeners revealed no significant main effect of EIT on perceptual responses in the AEM task ($\beta = 0.331$, $p = .127$). However, notable interactions between EIT proficiency scores and social guise conditions emerged, specifically with the American guise (`EIT` × `American guise`: $\beta = -1.132$, $p < .001$). Moreover, the three-way interaction between VOT step, EIT, and American guise condition was significant (`vot_step` × `EIT` × `American guise`: $\beta = 0.177$, $p < .001$). This finding indicates that proficiency effects were particularly pronounced when Chinese listeners perceived speakers as American, highlighting that proficiency played a contextually specific role (see Table 5.3 and Figure 5.4).

Table 5.3: Significant fixed effects for Chinese listeners in EIT vs. AEM. Random effects indicated substantial variability by `participant` ($SD = 4.63$) and a small effect by `POA` ($SD = 1.63$), with residual variance ($SD = 41.05$) reflecting within-subject trial noise.

Term	Estimate	SE	Statistic	p-value
(Intercept)	-47.85	20.53	-2.33	.021
`vot_step`	13.76	3.13	4.40	< .001
`guiseAmerican`	119.49	24.07	4.96	< .001
`guiseMandarin`	62.76	24.07	2.61	.009
`vot_step` : `guiseAmerican`	-18.62	4.27	-4.36	< .001
`vot_step` : `guiseMandarin`	-9.53	4.28	-2.23	.026
`EIT` : `guiseAmerican`	-1.13	0.25	-4.48	< .001
`vot_step` : `EIT` : `guiseAmerican`	0.18	0.05	3.94	< .001

Russian Listeners. For Russian listeners, analysis of the AEM task showed no significant main effect of EIT ($\beta = -0.007$, $p = .957$) nor significant interactions involving EIT and guise conditions (all p-values $> .05$). These results suggest that proficiency did not meaningfully impact Russian listeners' perceptual categorization patterns. Thus, voicing categorization for this group appears primarily determined by phonetic properties and social context, independent of proficiency differences (see Table 5.4 and Figure 5.4).

Table 5.4: Significant fixed effects for Russian listeners in EIT vs. AEM.

Term	Estimate	Std. Error	Statistic	p-value
`vot_step`	10.89	1.78	6.12	< .001
`vot_step` : `guisebaseline`	-6.00	2.55	-2.35	.019
`vot_step` : `EIT` : `guisebaseline`	0.06	0.03	2.38	.017

Overall, EIT performance did not exert a main effect on AEM voiceless categorization in either group, appearing only in interaction with guise or VOT step. Among Chinese listeners, EIT interacted significantly with the American guise ($p < .001$), such that higher-proficiency listeners showed more voiceless responses at earlier VOT steps, reflecting a shift toward an English-like perceptual boundary. For Russian listeners, proficiency had minimal impact; VOT step and social guise explained the majority of categorization patterns. Across all participants in the AEM task, the VOT step remained the strongest and most consistent predictor of voiceless categorization across all participants.

EIT and Deliberative Speech Categorization: Proficiency Effects in the VAS Task

I next examined whether participants' oral English proficiency, as measured by the Elicited Imitation Task (EIT), predicted their voiceless categorization patterns on the Visual Analogue Scale (VAS) task. Unlike the AEM dataset, the VAS dataset contains a larger sample ($N = 80$), as all participants who completed AEM also completed VAS, but not vice versa. Of these 80 participants, 74 had usable EIT scores, including 39 Chinese listeners and 35 Russian listeners.

The histograms in Figure 5.5 display the EIT score distributions by listener group. Chinese listeners showed a relatively symmetric, centralized distribution ($Mean =$ 95.39, $Median =$ 98, $Min =$ 57, $Max =$ 117), while Russian listeners' scores were more right-skewed, with a higher median and broader representation in the upper range ($Mean =$ 99.48, $Median =$ 105, $Min =$ 60, $Max =$ 120). Unlike the AEM subset, the Russian group here includes several lower-scoring individuals, resulting in a more even spread across the proficiency range.

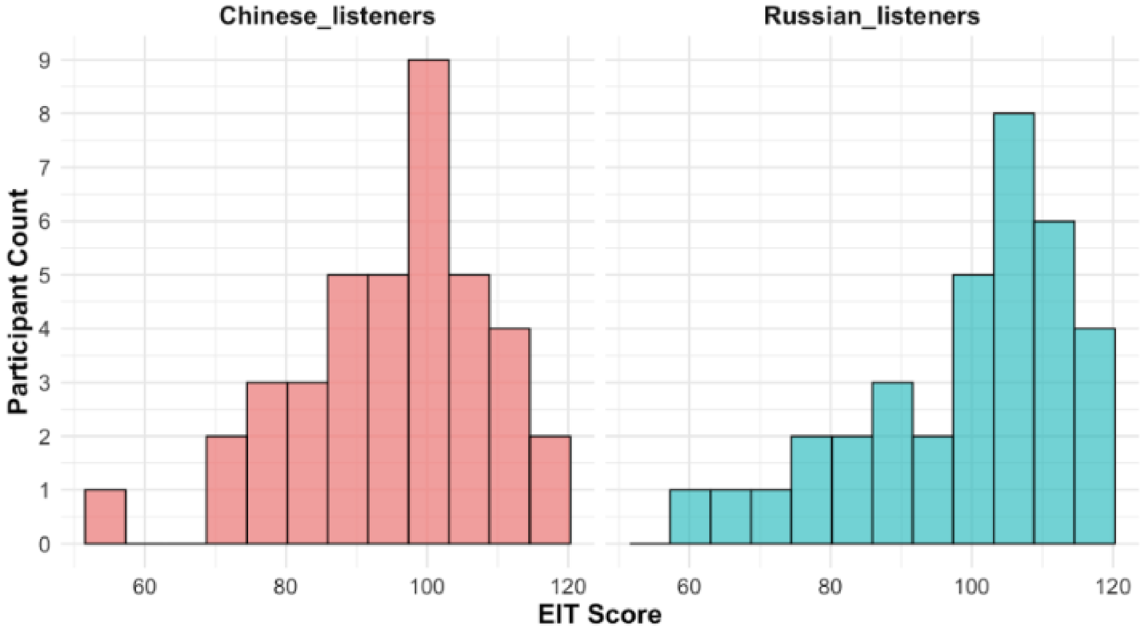


Figure 5.5: Distribution of unique EIT scores by listener group. Left: Chinese listeners; right: Russian listeners. Note: Bin width = 5 points.

To visualize how proficiency might relate to perceptual performance, EIT scores were divided into tertiles based on quantiles computed from this dataset. The breakpoints were:

- **Low proficiency**: EIT ≤ 92
- **Mid proficiency**: $92 <$ EIT ≤ 106
- **High proficiency**: EIT > 106

These tertiles were used to generate Figure 5.6, which plots average voiceless categorization responses across nine VOT steps by EIT tier, listener group, and guise condition. The figure reveals that EIT proficiency had minimal effect on categorization behavior in the VAS task.

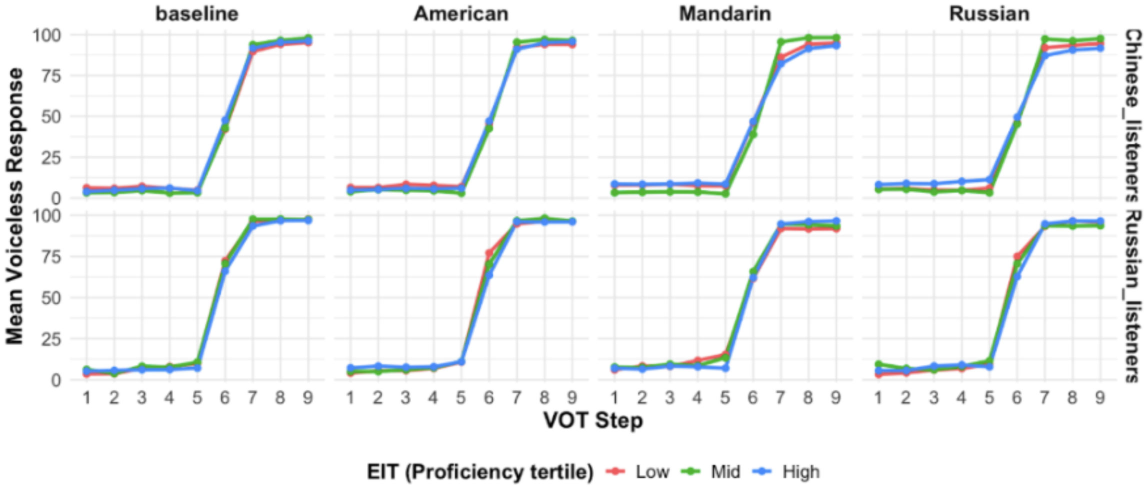


Figure 5.6: VAS voicing categorization by VOT step, EIT (proficiency tertile), guise, and group. Top row: Chinese listeners; bottom row: Russian listeners. Left-to-right panels represent the four experimental conditions: baseline, American, Mandarin, and Russian guise.

Unlike in the AEM data, EIT proficiency appears to exert very limited influence on VAS categorization behavior. For both Chinese and Russian listeners, perceptual patterns across the VOT continuum are remarkably similar across proficiency tiers. One possible exception is observed among Russian listeners in the Russian guise condition, where the mid-proficiency group shows slightly steeper categorization slopes, though the difference is marginal. The lack of clear tier-based effects may be partly due to loss of sensitivity when converting continuous EIT scores into tertile categories, a limitation addressed in the modeling section that follows.

EIT vs. VAS: Chinese Listeners. The linear mixed-effects analysis for Chinese listeners revealed no significant main effect of EIT on categorization responses in the VAS task ($\beta = -0.036$, $p = .651$), nor any significant two-way interaction between EIT and either VOT or guise. However, EIT did emerge in one notable three-way interaction: the interaction between VOT step, EIT, and the Russian guise condition was significant ($\beta = -0.044$, $p = .014$), indicating that higher English proficiency was associated with a weaker relationship between VOT and voiceless categorization in the Russian guise. Additionally, two other interaction terms involving `guiseRussian` reached significance: a main effect ($\beta = -22.31$, $p = .021$) and an interaction with VOT ($\beta = 3.93$, $p = .022$), suggesting some sensitivity to social context. See all fixed effects and magnitudes in Table 5.5.

Despite these findings, the overall influence of EIT on VAS performance appears relatively limited and context-specific. Importantly, in contrast to the AEM task—where EIT effects were clearest in the American guise—the Russian guise condition was the only context in which proficiency influenced Chinese listeners' VAS judgments.

Table 5.5: Significant fixed effects for Chinese listeners in EIT vs. VAS.

Term	Estimate	Std. Error	t	p-value
(Intercept)	-28.26	8.40	-3.36	.001
`step`	13.26	1.22	10.89	$< .001$
`guiseRussian`	-22.31	9.68	-2.30	.021
`step : guiseRussian`	3.93	1.72	2.28	.022
`EIT : guiseRussian`	0.26	0.10	2.54	.011
`step : EIT : guiseRussian`	-0.044	0.018	-2.45	.014

Random effects revealed modest variability across participants ($SD = 3.00$) and a more notable influence of place of articulation ($SD = 6.02$), while within-subject variability remained substantial (residual $SD = 24.88$), consistent with the VAS task's trial-by-trial response noise.

EIT vs. VAS: Russian Listeners. The linear mixed-effects model for Russian listeners revealed no significant main effect of English proficiency (EIT) on VAS categorization responses ($\beta = 0.035$, $p = .684$), nor any significant interactions between

EIT and guise or VOT step. This result suggests that, for Russian listeners, proficiency did not play a meaningful role in shaping perceptual categorization judgments on the VAS task. Instead, response patterns appear to have been primarily driven by phonetic information and social guise, independent of individual differences in oral English proficiency.

Only one predictor emerged as a robust and consistent driver of categorization: VOT step ($\beta = 15.53$, $p < .001$), reinforcing the reliability of the task design and the salience of the acoustic continuum in listeners' responses. Random effects indicated moderate variability across participants ($SD = 4.47$) and by place of articulation ($SD = 6.16$), with residual variance ($SD = 26.03$) again reflecting trial-by-trial response variability typical of VAS tasks.

As no fixed effects involving EIT were statistically significant, no fixed effects table is included for this group.

Cross-Group Analysis: Does English Proficiency Override L1 Effects?

Previous sections demonstrated that English oral proficiency as measured by EIT sometimes modulates voicing perception within listener groups, but it remains unclear whether EIT accounts for categorization differences between Chinese and Russian bilinguals. To address this question, this section models EIT effects across both groups together, with parallel models run for the VAS and AEM tasks. This approach determines whether English proficiency drives shared patterns of categorization or if group differences persist even after accounting for EIT. In doing so, the analysis directly addresses the core question: are bilinguals' perceptual boundaries more influenced by L1 phonology or by English proficiency in varied task contexts?

AEM: Significant Effects and Group Differences. The cross-group model revealed strong effects of both listener group and social guise on voiceless categorization. Across both groups, the presence of a social guise significantly increased voiceless responses relative to the baseline: voiceless categorization rose by 119.7 points in the American guise ($p < .001$) and by 62.9 points in the non-native guise ($p = .008$). Russian listeners showed significantly higher voiceless categorization responses than Chinese listeners ($\beta = 61.8$, $p = .015$), aligning with their L1's shorter positive VOT boundary and consistent with expectations.

Notably, English oral proficiency (EIT) did not exhibit a significant main effect on categorization ($\beta = 0.34$, $p = .139$), suggesting that English oral proficiency alone does not drive global shifts in categorization patterns. However, EIT did interact meaningfully with the selective listener group and social guise conditions.

Among Russian listeners, higher EIT was associated with lower voiceless categorization in the American ($\beta = -1.14$, $p < .001$) and matching non-native guise (Russian) conditions ($\beta = -0.50$, $p = .045$), suggesting that greater English proficiency may encourage stronger reliance on fine-grained phonetic cues or reduced reliance on social expectations in ambiguous cases. For Chinese listeners, EIT did not significantly predict voicing categorization in any guise condition, indicating that English proficiency exerted a weaker or nonsystematic influence on their perceptual boundaries. Additionally, a significant `EIT` × `group` interaction ($\beta = -0.62$, $p = .019$) confirmed that the effect of proficiency was not uniform across groups.

Higher-order interactions further reveal this nuance: three-way interactions involving EIT, VOT step, and guise were significant for both listener groups (e.g., `VOT` × `EIT` × `American guise`: $\beta = 0.18$, $p < .001$). This result indicates that proficiency effects were most pronounced in how listeners resolved ambiguous VOT cues in socially charged conditions. Notably, this interaction was strongest among Russian listeners

(`EIT` × `guise` × `group`: all $p < .001$), supporting the idea that English proficiency modulated their reliance on VOT under socially informative conditions.

In sum, these results suggest that listeners' L1 background continues to exert a strong and independent effect on categorization, even after accounting for English proficiency. However, EIT plays a meaningful role when social information about the talker is available, particularly among Russian listeners. Figure 5.7 illustrates this pattern: in both the American and non-native L1 guise conditions, higher-proficiency Russian listeners (blue line) display more categorical voicing boundaries, suggesting greater certainty or perceptual resolution. In contrast, proficiency-related shifts were more muted among Chinese listeners.

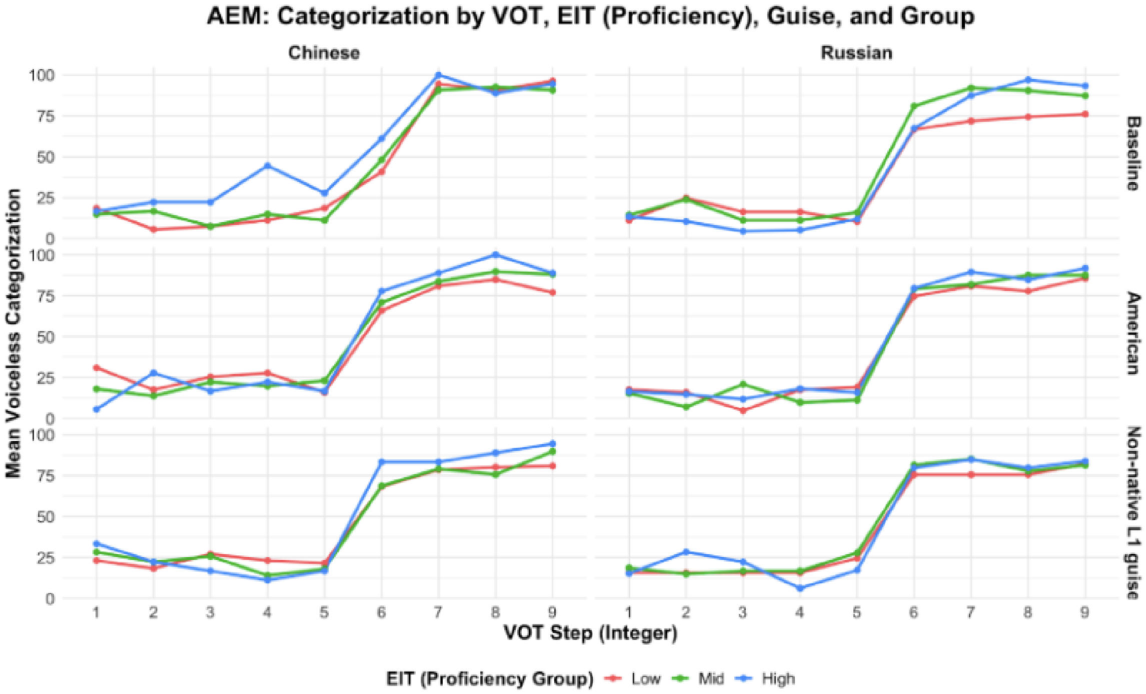


Figure 5.7: Mean voiceless categorization on AEM task by VOT, EIT, guise, and listener group.

Random effects analyses reveal only modest variability across listeners (SD = 5.75) and minimal influence of place of articulation (SD = 0.32), justifying their inclusion as random intercepts to account for any residual, subject-specific or articulatory variability.

VAS: Significant Effects and Group Differences. In the combined model for the VAS task, English oral proficiency (EIT) did not emerge as a significant overall predictor of voiceless categorization ($\beta = -0.036$, $p = .67$), and no significant group difference was observed between Chinese and Russian listeners ($\beta = -5.63$, $p = .63$). This effect contrasts with the AEM task, where listener group and EIT interacted more robustly, suggesting that VAS ratings are less sensitive to either L1 background or proficiency effects.

Instead, significant variation in VAS ratings stemmed from responses to the Russian guise, which elicited lower voiceless categorization overall ($\beta = -22.31$, $p = .024$), particularly in interaction with VOT step ($\beta = 3.93$, $p = .026$). A significant three-way interaction between EIT, VOT step, and Russian guise ($\beta = -0.044$, $p = .016$) further indicates that proficiency-related effects emerge only when specific social cues are present. In this case, listeners with higher EIT scores showed slightly less voiceless categorization in the Russian guise as VOT increased, suggesting that Russian guise was most salient for both listener groups and increased perceptual caution. Full model results are presented in Table H.1 in Appendix H.

Random effects estimates show that individual differences among listeners remain moderate ($SD = 3.77$), while place of articulation continues to have a small influence ($SD = 6.04$). These results suggest that, in this combined model, differences in voiceless categorization are shaped less by English proficiency or L1 group membership alone, and more by the specific interaction of social guise and the phonetic properties of the stimulus.

5.3.3 Results

Across all six analyses, results showed that L1 background exerted the most consistent influence, while proficiency only affected behavioral responses under specific conditions.

In the AEM task, which captures rapid eye movements in response to auditory stimuli, English oral proficiency influenced responses primarily through interactions with social guise. For Chinese listeners, higher proficiency was linked to more English-like voiceless categorization when the speaker was framed as American. For Russian listeners, proficiency correlated with more categorical responses, especially when social cues were present. These effects suggest that when listeners engage with speech in a more automatic, time-sensitive way, higher proficiency may help resolve ambiguity more decisively, and towards the directions driven by social expectations. Further, Russian listeners consistently showed more voiceless responses than Chinese listeners, reflecting their L1's phonological structure and suggesting that native language remains a key influence.

In contrast, the VAS task involved explicit judgments using a rating scale. Here, proficiency had very limited influence, surfacing only once in a higher-order three-way interaction with both social guise and acoustic cue. Group differences by L1 were also weaker. The more deliberate nature of the VAS task may reduce the weight of subtle proficiency-related processing differences, making responses more uniform across listeners.

Across both tasks, voicing decisions were most consistently predicted by the acoustic VOT cue itself. Yet when social information was introduced, and especially when perception was tracked in real time, English proficiency began to play a role. These findings point to a layered system, where L1 phonological structure shapes core perceptual boundaries, while oral proficiency selectively modulates these boundaries under conditions of increased social or cognitive demand. In short, bilingual listeners do not simply become more “native-like” as their English improves. Instead, their perceptual systems remain fundamentally shaped by their first language, with proficiency effects emerging only under certain social or cognitive conditions. This suggests that some differences that were initially attributed to “being Chinese vs. Russian” in earlier chapters can actually be explained by how proficient each listener is in English, a factor which varies within groups. In other words, proficiency cuts across the L1 groups to influence socially modulated speech perception.

The remaining variability even after accounting for proficiency motivates our final set of analyses: can cognitive style assessed via Autism Spectrum Quotient further explain who is most vs. least adaptable to social cues?

5.4 Sociocognitive Styles as Reflected in Autism Quotient Spectrum

Recent research highlights the importance of accounting for individual differences in language processing, particularly in tasks involving ambiguity, perceptual flexibility, and social cue integration. The Autism Spectrum Quotient (AQ) developed by Baron-Cohen et al. (2001) is a widely-used self-report measure for adults that quantifies autistic-like traits in neurotypical populations (Baron-Cohen et al., 2001; Baker et al., 2025). The AQ comprises five subdomains: Social Skills, Attention Switching, Attention to Detail, Communication, and Imagination. Previous work has linked variation in these traits to differences in both perceptual and social processing styles (Stewart and Ota, 2008; Yu, 2010; Yu et al., 2011).

Given this dissertation’s focus on how bilingual listeners perceive ambiguous speech in the presence of social cues, I hypothesize that AQ, both as an overall score and at the subscale level, may help explain variability in listeners' responses, above and beyond L1 background or experimental condition. For example, individuals with higher scores on certain AQ traits might be less susceptible to changes across experimental conditions at the same VOT step, possibly due to a greater tendency to focus on acoustic details rather than social context.

5.4.1 Administration and Data Handling

The AQ questionnaire was administered via the online platform NovoPsych (NovoPsych, 2025) to all participants following the AEM experiment. Participants completed the test in the Linguistics Lab using a computer and were provided with official translations in their native languages—Taiwanese Chinese version (Gau, 2024); Russian version (Shabalin, 2024)—as published by the Autism Research Centre. The assessment consists of 50 items and typically requires five to ten minutes to complete. Upon completion, the platform automatically generated a report for each participant, summarizing the raw score, percentile (relative to autistic and neurotypical populations), and a categorical descriptor (not consistent, consistent, or pronounced) indicating the degree of alignment with autistic traits.

Importantly, the test algorithm adjusts for gender and age norms when converting raw scores into percentile rankings and categorical descriptors. As a result, the same raw score near a classification threshold may be interpreted differently depending on the test-taker's gender or age group. For example, a raw score at the border between categories may be classified as "consistent with the autistic population" for a female participant, but as "not consistent" for a male participant, reflecting known differences in AQ score distributions across demographic groups.

A total of 42 listeners completed the AEM task and were invited to complete the AQ. After exclusion of cases with incomplete AQ data, the final sample comprised

35 participants (17 Chinese and 18 Russian listeners). Data exclusions were primarily due to link expiration or incomplete test submission.

5.4.2 Analytical Approach

Group Differences in Overall AQ Score Distribution

To explore potential group-level differences in cognitive style, I first visualized the distribution of overall AQ percentiles and categorical classifications for each listener group (see Figure 5.8). Among Chinese listeners, one participant was classified as "pronounced" autistic and three as "consistent" with autistic norms, while the rest (N) were "not consistent." Russian listeners were even less aligned with autistic norms: only one was "consistent," and the rest ($N = 16$) were "not consistent," clustering primarily at the lowest end of the AQ percentile scale (most scoring below the 25th percentile).

These distributions are meaningful for current predictions: if AQ captures stable individual differences in social and phonetic processing, the greater alignment of Chinese listeners with autistic norms may predict less pronounced or flexible shifts in perceptual boundary in response to social cues, while the lower AQ scores among Russian listeners may be associated with greater perceptual flexibility across guises. This hypothesis is evaluated in the following analyses, which test whether group-level differences in AQ distribution translate into differences in how listeners adapt to talker guise across the AEM task. A full breakdown of AQ percentiles by subscale and listener group is provided in Figure J.1 in Appendix J.

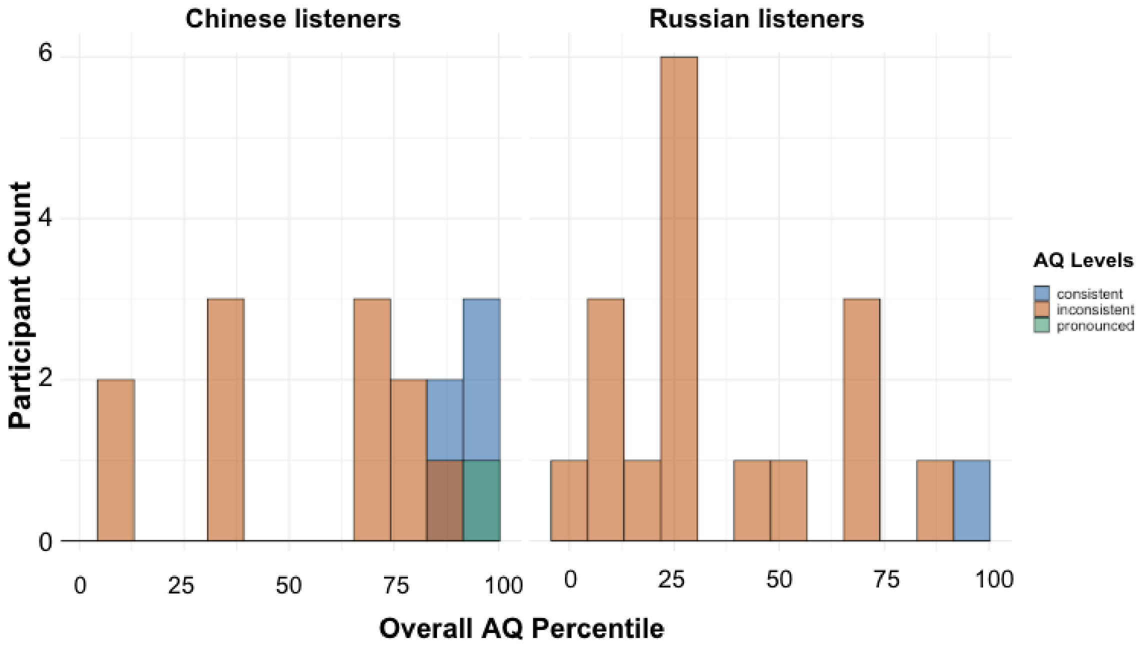


Figure 5.8: Distribution of overall Autism Quotient (AQ) percentiles and categorical levels by listener group.

Effects of AQ Subscale on Individual Shifts

To identify which aspects of autistic-like traits are most predictive of perceptual boundary shifts, I adopted a systematic model comparison approach. Each analysis used a mixed-effects regression model with the following structure:

```
response ∼ aq_metric * guise + vot_step + (1 | ID) + (1 | poa)
```

Here, `response` is participants' voiceless categorization along a VOT continuum, and `aq_metric` refers to the overall AQ score or any subscale (Social Skill, Attention Switching, Attention to Detail, Communication, Imagination). The model included all AQ metrics and their interactions with `guise` (`baseline, native, non-native`), with `vot_step` as a main effect and `participant ID` and `POA` place of articulation as random intercepts. This approach allows me to directly test whether AQ traits predict listeners' sensitivity to social guise above and beyond language background or other factors.

All six AQ metrics were initially entered as predictors. At each stage, I removed the least predictive metric (i.e., defined by the highest p-value) and refitted the model. This stepwise process continued until only AQ metrics with significant main effects or interactions remained. At every step, model fit was evaluated using AIC/BIC and residual variance, ensuring that the final model was as simple as possible while still providing explanatory power. This model selection strategy allowed me to identify the cognitive dimensions most relevant for predicting whether, and how, listeners shift their category boundaries in response to social cues.

The best-fitting model, as indicated by the lowest AIC/BIC, included Communication, Attention Switching, Imagination, and overall AQ (with Attention to Detail and Social Skill dropped):

```
response  ~  aq_all * guise + aq_attention_switching * guise +
             aq_communication * guise + imagination * guise +
             vot_step + (1 | ID) + (1 | poa)
```

The best-fitting model revealed that only the interaction between Communication and the non-native guise condition reached significance ($p = 0.0485$, $\beta = -0.100$). This negative estimate indicates that, as autistic-like communication traits increase, listeners are less likely to increase their proportion of voiceless categorizations when presented with a non-native talker. In other words, individuals with more pronounced communication-related autistic traits were less likely to shift their perceptual boundary toward voicelessness in the presence of a socially cued talker, particularly when that talker matched their own linguistic background.

As shown in Figure 5.9, in the baseline condition, the higher the alignment in communication with autistic populations is associated with a slightly enhanced percentage of voiceless categorization. In contrast, in the presence of a social guise, especially a non-native condition, this pattern was reversed, where increased communication corresponded with weakened voiceless categorization in the AEM task. Within the experimental conditions, the effect is especially pronounced for non-native guise.

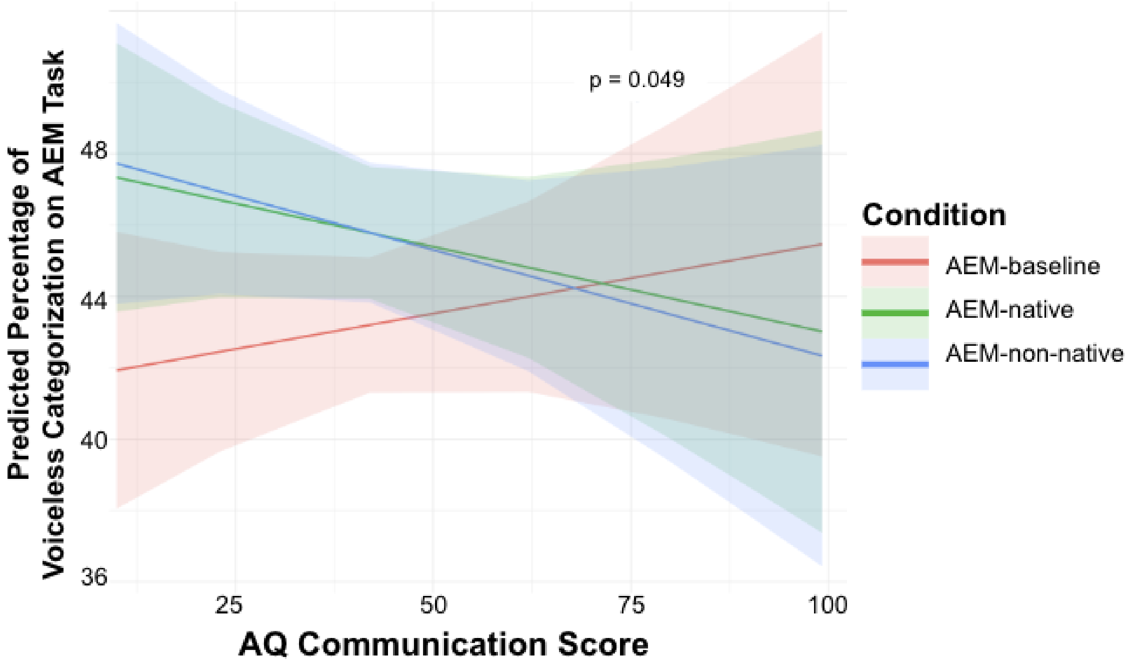


Figure 5.9: Effects of individual communication traits on voiceless perception across talker conditions.

Correlation between Individuals' AQ Metrics and Shift from Baseline

To complement the earlier mixed-effects modeling, I visualized and tested the relationship between individual AQ metrics and the degree to which each participant shifted their perceptual boundary in response to talker guise. This analysis serves two purposes. First, it allows for inspection of effects at the individual level, moving beyond aggregate trends to see how AQ traits might modulate perception differently across listeners. Second, by modeling “shift from baseline” separately within voiced (steps 1–5) and voiceless zones (steps 6–9), I account for the fact that social guises could plausibly affect listeners in opposite directions depending on their L1 phonological boundary (i.e., short-lag for Russian, long-lag for Mandarin). This distinction is crucial, as any genuine effect of social guise would be expected to push listeners' categorization in different directions across zones, rather than uniformly across the VOT continuum.

For each participant, the mean shift is calculated from baseline in both native and non-native guises, separately for the voiced and voiceless zones. Figure 5.10 illustrates the relationship between overall AQ percentile and this shift from baseline, with trends and confidence bands plotted by talker guise, voicing zone, and listener group. Each panel displays both group-level patterns (via fitted lines) and the full spread of individual responses (scatterplot).

Substantial variability exists across participants, but group differences are clearest in the voiceless zone. As AQ percentile increases, Mandarin listeners tend to show reduced voiceless categorization in both guises, consistent with their L1 Mandarin phonological boundary. Russian listeners, by contrast, exhibit increased voiceless categorization with higher AQ scores, also reflecting their native Russian cue boundary. These trends are present regardless of whether the guise matches the listener's native language.

While there is no strong linear association across all participants, these group-level patterns suggest that higher AQ may anchor perception more closely to L1 norms, while lower AQ listeners show greater flexibility, potentially shifting toward L2-like responses when exposed to social cues.

In the voiced zone, there is also considerable individual variability, but the two listener groups show much greater overlap in both the magnitude and direction of shift from baseline. Here, shifts tend to cluster around zero or show small downward trends. It is important to note that the gaze-based metric (`first_looked_left`) used here reflects only whether the first off-centered look was leftward (voiceless categorization), so a downward shift in this panel indicates a reduced percentage of voiceless categorization, but does not necessarily mean participants confidently categorized the stimulus as voiced. Some listeners may simply have kept their gaze centered, resulting in fewer "voiceless" looks without a clear voiced identification.

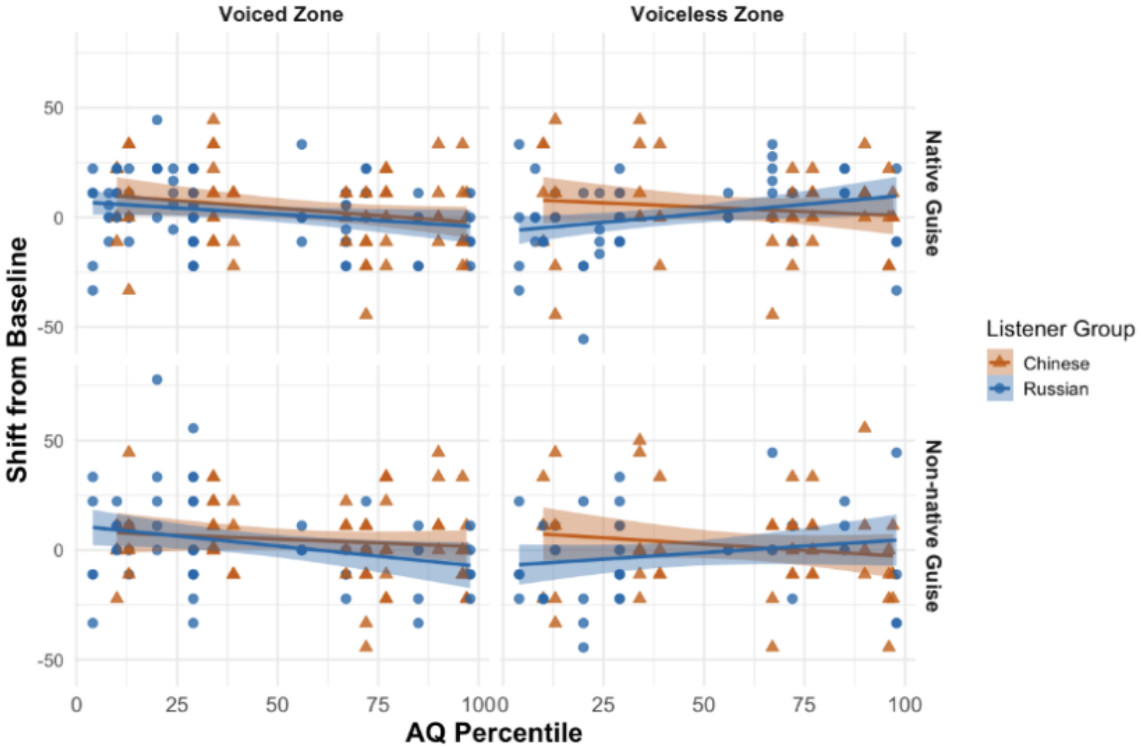


Figure 5.10: Relationship between AQ percentile and shift from baseline by condition, zone, and listener group (AEM task).

Expanding on the previous analysis, Figure 5.11 presents the same plots for each AQ subscale, by condition and voicing zone. These visualizations confirm the patterns seen for overall AQ: if a particular trait had a strong, systematic influence on perceptual shift, one would expect to see clear sloping trends across the panels. In practice, the directionality observed for `AQ_all` is largely replicated in the subscale plots. Some subscales, such as Attention to Detail and Social Skill, show nearly flat, overlapping lines for both groups, indicating little or no effect. In contrast, Imagination and Communication subscales show more pronounced group-level differences, echoing the patterns observed for overall AQ, with greater separation between Russian and Chinese listeners as AQ increases.

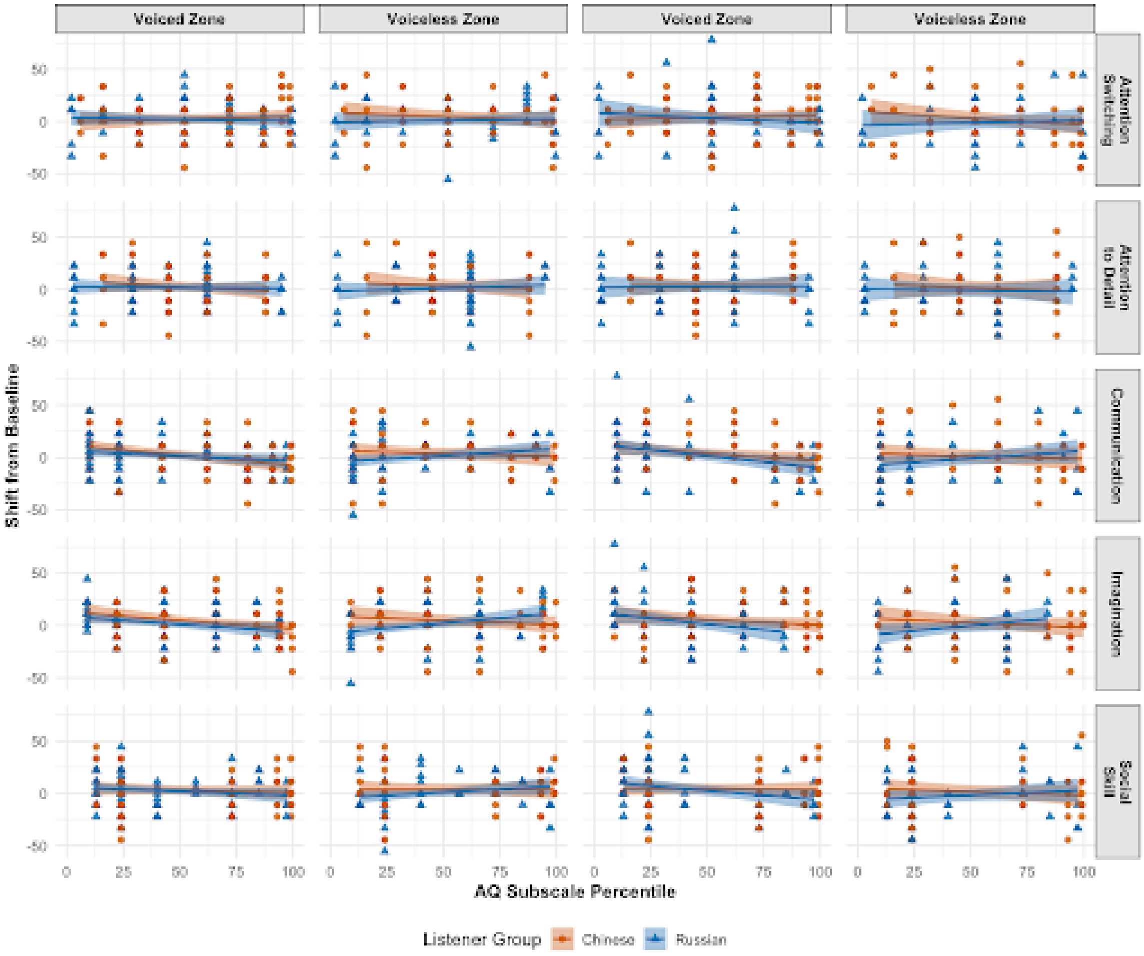


Figure 5.11: AQ subscale effects on VOT shift by experimental condition and voicing zone.

To formally assess these patterns, Pearson's correlation analyses were conducted between each AQ metric and the magnitude of shift from baseline in the voiceless zone, where group and guise effects were most pronounced. Correlations were run separately for native and non-native guises and for each listener group, resulting in 24 tests (6 AQ metrics × 2 guises × 2 groups). The analysis is limited to the voiceless zone because, as discussed above, the binary leftward gaze metric more reliably captures voiceless categorization, whereas decreased voiceless responses in the voiced zone do not necessarily entail confident voiced categorization. None of these correlations were statistically significant, indicating that AQ traits do not consistently predict individual shifts in categorization in this context. Results are interpreted in the section below.

5.4.3 Results

Exploration of AQ's influence on responses in the AEM task revealed that individual differences on the Communication subscale play a selective but meaningful role in modulating bilingual listeners' use of social cues in speech perception. This relationship emerged as the only significant effect among the five AQ subscales, linking listener's communication-related tendencies to their categorization of voicelessness as captured by the binary leftward gaze metric in the AEM paradigm.

Follow-up mixed-effects modeling confirmed this pattern: Communication was the only AQ subscales to show a significant interaction with talker guise, specifically in the non-native condition. The negative estimate for this interaction indicates that listeners with higher autism-aligned communication traits are less likely to increase voiceless categorization when exposed to a non-native guise. This suggests that stronger autism-aligned communication tendencies may buffer or dampen the influence of social context on speech perception, potentially reflecting lower sensitivity to the social markedness of non-native speakers.

Visualization of individual-level shifts clarified these trends. When overall AQ scores were low, both Russian and Mandarin listeners tended to show their L2 English-like shifts in perceptual boundary or no shift at all, regardless of guise. As AQ scores increased, both groups exhibited patterns more aligned with their L1-specific boundaries: Chinese listeners reduced voiceless categorization, while Russian listeners increased it, regardless of whether the social guise matched their native background. This convergence toward L1-like perception among high-AQ individuals suggests that autistic-like communication traits may anchor listeners more firmly to native phonological norms, overriding flexible, L2-like adaptation to social cues.

Importantly, the substantial variability observed in individual perceptual shifts is not simply random noise. Rather, the consistency of these trends by AQ and listener group, particularly in the voiceless zone, suggests that this variability is systematic rather than random. The patterns observed are not simply noise; instead, they reflect meaningful, individual-level differences in how cognitive style and L1 background jointly influence listeners' sensitivity to social cues in speech perception.

Finally, Pearson's correlation analyses across all AQ metrics, conditions, and groups in the voiceless zone (24 tests in total) found no statistically significant associations between AQ and perceptual shift magnitude. While some trends approached significance, none were robust or consistent. These results highlight the complexity of individual variability in social cue integration, indicating that while Communication plays a group-level role, there is not yet a statistically significant, direct correlation between AQ traits and perceptual flexibility in this context. These cognitive style effects that are captured, while subtle, highlight that even within listeners of the same L1 background and proficiency, personal cognitive traits can modulate speech perception. This observation adds a new dimension to our understanding of the perceptual flexibility observed in Chapters 3 and 4.

5.5 Discussion

This chapter adopted an explicitly individual-differences approach to examine how L2 oral proficiency (EIT) and cognitive style (AQ) modulate socially guided voicing perception in Russian-English and Mandarin-English bilinguals. Moving beyond essentialist group comparisons, I modeled how these factors shape perceptual boundary placement and flexibility across two tasks (implicit AEM vs. explicit VAS). Our analyses reveal that while L1 background exerts a persistent influence, individual variability in proficiency and cognitive traits systematically conditions listeners' sensitivity to social and acoustic cues.

5.5.1 Summary of Key Findings and Alignment with Predictions

This section reports key summaries on key findings and their alignments with the predictions.

RQ1: Task Effects (AEM > VAS)

As predicted, social guise effects were more robust in the AEM task than in the VAS task. Across both listener groups, gaze-based AEM responses revealed more gradient and dynamic shifts in perceptual boundaries, whereas VAS ratings tended to be more categorical and self-consistent. This pattern was especially pronounced for Chinese listeners, who showed stronger and more consistent modulation in gaze behavior under social guise, supporting the view that implicit measures like eye-tracking better capture genuine perceptual adjustments by minimizing self-monitoring and social desirability biases (Fisher, 1993; Hay and Drager, 2010). For Russian listeners, the AEM task revealed complex, bidirectional shifts, sometimes amplifying or dampening social-cue effects depending on the guise and acoustic context, indicating that social cue integration is not uniform, but depends on both task and listener group.

RQ2: Proficiency (EIT) as Contextual Moderator

Higher EIT scores predicted more English-like voicing boundaries in the AEM task, but only under certain social contexts. Among Chinese listeners, greater proficiency facilitated more native-like categorization, especially when the talker was framed as American. For Russian listeners, proficiency was associated with sharper, more categorical responses in the presence of social cues. These results partially support the prediction that proficiency promotes L2-aligned perception (Flege, 1995), but also reveal that this effect emerges primarily when social cues activate L2 schema. In contrast, EIT scores had minimal impact on VAS responses, likely because the explicit, reflective nature of the rating task allowed listeners time to filter their judgments, and the use of syllabic stimuli encouraged more categorical rather than gradient decisions. Notably, Russian listeners showed more voiceless responses overall than Chinese listeners, reflecting persistent influence of L1 phonology even when proficiency and social context are accounted for.

RQ3: Cognitive Style (AQ) Anchors L1 Influence

In partial support with with earlier prediction, one autistic-like trait Communication modulated social cue sensitivity at the group level: higher Communication scores (more “autistic”-aligning) predicted reduced shifts toward voiceless categorization in response to both social guises, but particularly the non-native guises, reflecting a greater reliance on bottom-up processing, or default to L1-aligning processing style (Yu, 2010). Notably, listeners with higher AQ scores tended to converge on L1-specific boundaries—Russians increasing voiceless responses, Chinese decreasing them—suggesting that cognitive style reinforces native phonological anchoring when social demands are high. However, when individual-level Pearson correlations were examined across all AQ subscales, experimental conditions, and listener groups, no significant associations emerged, highlighting that while AQ-Communication shapes group-level trends, individual relationships with perceptual flexibility are complex and not consistently robust. Other AQ subscales did not emerge to be significant.

5.5.2 High-Level Takeaways: What Individual Assessments Reveal

Across tasks and analyses, oral English proficiency, as measured by the EIT, emerged as the strongest perceptual moderator, facilitating L2-like re-tuning primarily when social cues, such as the American guise, heightened the salience of English. This effect was evident only in the AEM task, highlighting that proficiency's influence is tied to rapid, automatic processing rather than explicit reflection. In contrast, proficiency had minimal impact on VAS responses, likely because the explicit, categorical nature of the rating task encourages reliance on established phonological categories rather than dynamic adjustment.

Cognitive style, measured by the Autism Spectrum Quotient, played a more selective role. While global AQ scores showed limited predictive power, the Communication subscale was significant in tempering social cue integration: listeners with higher Communication scores demonstrated reduced flexibility in voicing categorization under non-native guises, reflecting a greater anchoring to bottom-up acoustic cues. This subscale, which often translates into individuals' pragmatic awareness, proved more explanatory than overall AQ for individual variability in social adaptation. However, neither proficiency nor cognitive style fully attenuated the influence of L1 background. Russian listeners consistently exhibited higher rates of voiceless responses, in line with their native short-lag VOT boundaries, whereas Chinese listeners' social-cue sensitivity was modulated by both proficiency and AQ subscales.

These individual differences help explain why group-level patterns diverged across experiments. For Russian listeners, proficiency-mediated flexibility was evident: those with higher EIT scores showed steeper, more categorical boundaries when social cues activated L1/L2 conflict. For Chinese listeners, high proficiency enabled more English-like boundaries under American guise, while lower AQ scores predicted greater social cue integration, though overall AQ distributions were comparable between groups. Notably, at high oral English proficiency levels, both groups converged in exhibiting more decisive, cue-based categorization when social cues were present. This convergence suggests that, despite persistent differences in native boundary placement, proficiency enables both Russian and Chinese listeners to adopt similar processing strategies for rapid resolution when confronted with social information.

5.5.3 Methodological Reflections and Future Directions

While the analyses in this chapter robustly addressed the core research questions, two methodological limitations merit discussion.

First, our analysis of EIT in this study (by including EIT as an interaction term with task, guise, and VOT step) only indirectly addressed how proficiency relates to context-driven perceptual shifts. A more illuminating approach would be to calculate each listener's shift in category boundary from the baseline to each guise condition, and then directly correlate those continuous shift scores with proficiency measures like EIT or oral proficiency. This method would capture the gradient, individual nature of bilingual perceptual flexibility far better than group-level comparisons. For example, it could reveal whether higher-proficiency listeners show larger or smaller shifts in their phonetic boundaries across social contexts – insights that might be missed with the broader modeling approach. Additionally, simplifying the statistical model could improve interpretability: VOT step's influence is largely systematic, so it could be treated as a main effect rather than entangled in high-order interactions. Focusing on the key factors (listener group, guise, and task) or using a stepwise model selection to remove non-predictive terms would likely yield a more parsimonious model.

Second, AQ effects were analyzed only within the AEM task, aligning with the theoretical focus on implicit, real-time social cue integration. However, extending AQ analyses to the VAS task could shed light on how cognitive style influences more reflective, self-monitored responses. Unlike AEM, VAS allows time for participants to consciously evaluate and potentially reinterpret ambiguous speech signals in light of social information. This may be especially revealing for understanding how individuals with varying AQ profiles integrate social cues when they are fully aware of them. For instance, listeners with low AQ scores, often associated with heightened social sensitivity, might display stronger alignment with socially congruent interpretations, even when the acoustic evidence is ambiguous. In contrast, those with higher AQ scores may either ignore the guise manipulation or resist adjusting their responses, showing reduced flexibility. Thus, AQ-VAS analyses could uncover strategic, awareness-driven components of social modulation in speech perception that remain latent in more automatic tasks like AEM.

In sum, these findings contribute to a growing body of work recognizing bilingual listeners as dynamic agents rather than static exemplars of group identity. Rather than erasing L1 influence, proficiency and cognitive style dynamically reconfigure how social and acoustic cues are weighted in real time. Russian and Chinese listeners diverge in where they place their boundaries, yet proficiency enables both groups to make more decisive, cue-based decisions, and cognitive traits regulate the degree of social-cue integration. The exploratory analyses presented here reveal bilingual perception as a multidimensional, context-dependent process—marking a paradigm shift from group-centric to agent-centered models of language (Eckert, 2018). Chapter 6 will synthesize these insights into a unified framework for socially embedded bilingual speech processing.

Chapter 6

CONCLUSION

This dissertation set out to investigate how Mandarin-English and Russian-English bilingual listeners integrate social expectations with acoustic cues during speech perception. Building on the foundation established in Chapters 1 and 2, I examined three overarching research questions:

1. *Do bilingual listeners shift their voicing category boundaries when social information about a talker (e.g., their presumed L1 background) is manipulated?*
2. *Are such shifts observable both in explicit categorization tasks (Chapter 3) and in implicit, real-time processing measures (Chapter 4)?*
3. *How do individual listener characteristics—such as English oral proficiency and cognitive-social traits—modulate these effects (Chapter 5)?*

These questions were grounded in evidence that bilingual speech perception is inherently dynamic and context-sensitive. Prior research has shown that bilinguals can flexibly adjust their perceptual boundaries based on the linguistic environment, but less was known about how subtle social cues, such as a talker's perceived ethnicity or language background, shape this adaptation. Additionally, while many sociophonetic studies focus on group-level trends, I sought to explore individual variation in perceptual strategies, asking whether certain listener traits predict susceptibility to social-contextual modulation. The subsequent sections summarize the methodological approaches and key findings of each chapter, evaluate how these findings align with our predictions, discuss the study's limitations, and outline directions for future research.

6.1 Summary of Methodological Approaches and Findings

To address these questions, I implemented three empirical studies described in Chapters 3–5. Each experiment was designed to examine a different facet of bilingual sociophonetic perception. Each experiment probed a different level of speech perception: explicit categorization, implicit real-time processing, and individual variation. Below, I synthesize the methodology and main findings of each.

6.1.1 Visual Analog Scale (VAS) Task (Chapter 3)

The first experiment employed a matched-guise VAS task to measure explicit phonetic categorization. Participants listened to a nine-step VOT continuum spanning the boundary between voiced and voiceless stops (e.g., /ba/-/pa/, /da/-/ta/, /ga/-/ka/). The same set of acoustic stimuli was presented across four listening blocks: a baseline (with no social information) and three social guise conditions in which the talker was described as American, Russian, or Chinese. Before each block, participants viewed short videos and written descriptions introducing the supposed talker identity. After each stimulus, participants rated its category membership on a continuous slider, producing a gradient measure of perception sensitive to subtle boundary shifts. This design tested whether participants' phonetic boundaries would shift toward the expected VOT norms of the talker's portrayed L1.

The results revealed clear socially modulated boundary shifts. Under the Russian guise (short-lag voicing norms), participants categorized ambiguous tokens as voiceless at shorter VOTs, whereas under the Chinese guise (long-lag norms), they tended to categorize the same stimuli as voiced. The baseline boundary fell between these two extremes. These shifts were statistically significant, demonstrating that social expectations alone, without any acoustic change, can recalibrate phonetic categorization. The VAS results thus strongly supported the hypothesis that bilingual listeners dynamically adjust their perceptual boundaries based on talker identity.

6.1.2 Anticipatory Eye Movement (AEM) Task (Chapter 4)

The second experiment (Chapter 4) extended the investigation from explicit judgments to implicit, real-time processing, using a webcam-based Anticipatory Eye Movement (AEM) paradigm. Participants were trained to associate voiceless-initial words with a left visual target and voiced-initial words with a right target, and their earliest gaze shifts during ambiguous VOT trials provided a window into pre-conscious categorization decisions. As in Chapter 3, social guises were introduced through short video primes describing the talker's background, allowing us to test whether visual social cues could influence the earliest stages of auditory processing.

This study was methodologically successful in validating a low-cost, scalable eye-tracking approach for sociophonetic research. Even after filtering out sessions with low video quality, 87 of 93 usable sessions (93.55%) met the accuracy threshold for baseline categorization, confirming that the AEM task captured reliable, category-aligned gaze behavior.

Beyond validating the paradigm, the AEM data uncovered meaningful between-group differences: Chinese listeners exhibited significant and consistent talker-induced perceptual shifts that aligned with the talker's presumed L1, whereas Russian listeners showed little to no social modulation. This striking contrast (not captured in VAS!) prompted deeper reflection on differences between the groups in raciolinguistic ideologies, language security, and cultural communication styles. It also provided a natural motivation to investigate individual differences in Chapter 5, where incorporating English proficiency (EIT) scores revealed clear `guise` × `proficiency` interactions, and autistic traits in Communication produced patterned effects on voiceless categorization—all of which demonstrate AEM's sensitivity to individual-level variation and implicit biases.

Chapter 4 also presents a methodological innovation, as we developed a completely new data-processing pipeline—transforming OpenFace gaze outputs into trial-aligned measures and reliable screen locations—for which no existing tools or guidelines previously existed. This pipeline now serves as a blueprint for future sociophonetic and bilingual perception research using webcam-based AEM.

6.1.3 Individual Differences Analysis (Chapter 5)

The final component investigated how listener-specific traits moderated the effects observed in the previous experiments. Participants' English oral proficiency was assessed via an Elicited Imitation Task (EIT), providing an objective index of productive L2 ability. Additionally, participants completed the Autism Spectrum Quotient (AQ) questionnaire, which measured five subscales—Social Skills, Attention Switching, Attention to Detail, Communication, and Imagination—capturing individual differences in social cognition and attentional style. Statistical models tested whether proficiency or AQ scores predicted the magnitude of perceptual shifts on the VAS and AEM tasks. By combining explicit ratings, implicit gaze measures, and individual-difference metrics, the research design provided a multi-dimensional view of how bilinguals integrate social and acoustic information.

The analysis revealed that higher English proficiency was associated with more stable categorization, often resulting in smaller guise-induced shifts on the VAS task. Between the two tasks, English proficiency attenuated the guise effect more in the gradient AEM task than the VAS task. AQ scores also played a role in the "Communication" subscale at the group level, where higher AQ individuals displayed a shift to L2 English norms in the presence of any social guise, while low-AQ individuals aligned with their L1 norms under socially primed conditions.

6.2 Implications for a Sociophonetic Account of Bilingual Perception Processing

The findings support a context-sensitive model of speech perception, where top-down social expectations interact with bottom-up acoustic cues. Explicit shifts observed in Chapter 3 align with exemplar-based and adaptive models, suggesting that bilingual listeners flexibly re-weight acoustic dimensions based on perceived talker identity. The results also refine theories of bilingual perception by highlighting the role of "social mode switching"—listeners may toggle between L1- and L2-tuned boundaries depending on social context.

Moreover, Chapter 5 highlights that perception cannot be fully understood without individual difference parameters, such as language experience and social sensitivity. This point suggests that models like the Speech Learning Model (Flege et al., 1995) or the Perceptual Assimilation Model (Best, 1994) could be extended by incorporating social weighting factors that vary across individuals.

6.3 Limitations

Despite these findings, several limitations should be acknowledged. These limitations arise from certain experimental design choices, practical constraints, and the scope of the research.

- **Single-Voice Guise Manipulation.** Using one voice across all guises reduced ecological validity. Some participants recognized the voice was unchanged, yet still exhibited shifts, an intriguing finding that suggests conceptual framing alone can bias perception.

- **Within- vs. Between-Subject Design.** Due to accumulated social knowledge effects and fatigue, I analyzed only participants' first social block, reducing within-subject comparability. A shorter, two-block design (baseline + one social guise) would have preserved within-person contrasts and improved statistical power.

- **Webcam Eye-Tracking Constraints.** The AEM task's resolution was lower than lab-based systems, possibly obscuring subtle real-time effects.

This dissertation demonstrates that bilingual speech perception is adaptive and context-sensitive. Explicit tasks (Chapter 3) showed that even minimal social cues shift voicing boundaries, while implicit measures (Chapter 4) revealed subtler, less uniform effects. Individual differences (Chapter 5) highlighted that these shifts are shaped by proficiency and social-cognitive traits, such Communication. By addressing the research questions outlined in Chapter 1, I provide empirical evidence that bilingual listeners actively integrate social expectations with acoustic cues rather than relying on fixed phonetic boundaries. These findings refine existing models of speech perception by emphasizing dynamic social tuning and individual variability.

Although further work is needed to consolidate these insights into a formal theoretical framework, this research lays important groundwork for understanding how identity, cognition, and experience jointly shape bilingual speech perception. Perception is not simply an acoustic process but a socially and cognitively mediated act of interpretation. I hope these findings inspire further exploration of how bilinguals navigate this complex interplay of sound and society.

6.4 Future Directions

Building on these findings and limitations, future research can take several directions:

- **Refined Experimental Designs.** Employ a more efficient VAS paradigm involving just two blocks—Baseline and one Social guise—randomized across participants. This adjustment would reduce fatigue and learning effects while allowing direct within-subject comparisons. Counterbalancing social contexts across participants would still enable multi-guise analysis. Using distinct talkers for each guise, alongside a classic matched-guise setup, would improve credibility and test generalizability.

- **Integrating Rich Social Data.** Future studies should include measures of socio-cultural identification and social networks to explore whether bilinguals with more diverse interactions or stronger cultural ties show greater flexibility in shifting phonetic boundaries. Social network analysis could quantify exposure to various accents and link it to susceptibility to social cues. Other demographic factors such as length of stay in the USA, age of arrival, could all contribute meaningfully in explaining between listener differences. I had collected these data via the use of LEAP-Q questionnaire, and can use this data for future analyses.

- **Linking Perception and Production.** The relationship between perception and production warrants deeper exploration. Follow-up studies could compare participants' VOT shifts with their production data. Do those who produce intermediate VOTs also show smaller perceptual shifts? Are "high adapters" consistent across both perception and production?

- **Analyzing Speech Across Contexts.** Speech from interviews and EIT recordings could be analyzed alongside controlled reading tasks to investigate how voicing boundaries vary across awareness levels. Comparing perception and production across these contexts would reveal how social and cognitive factors interact across modalities.

- **Expanding Populations and Phonemic Contrasts.** Extending this work to other language pairs (e.g., tonal contrasts or vowel systems) and including monolinguals could clarify which effects stem from bilingualism versus general sociolinguistic processes.

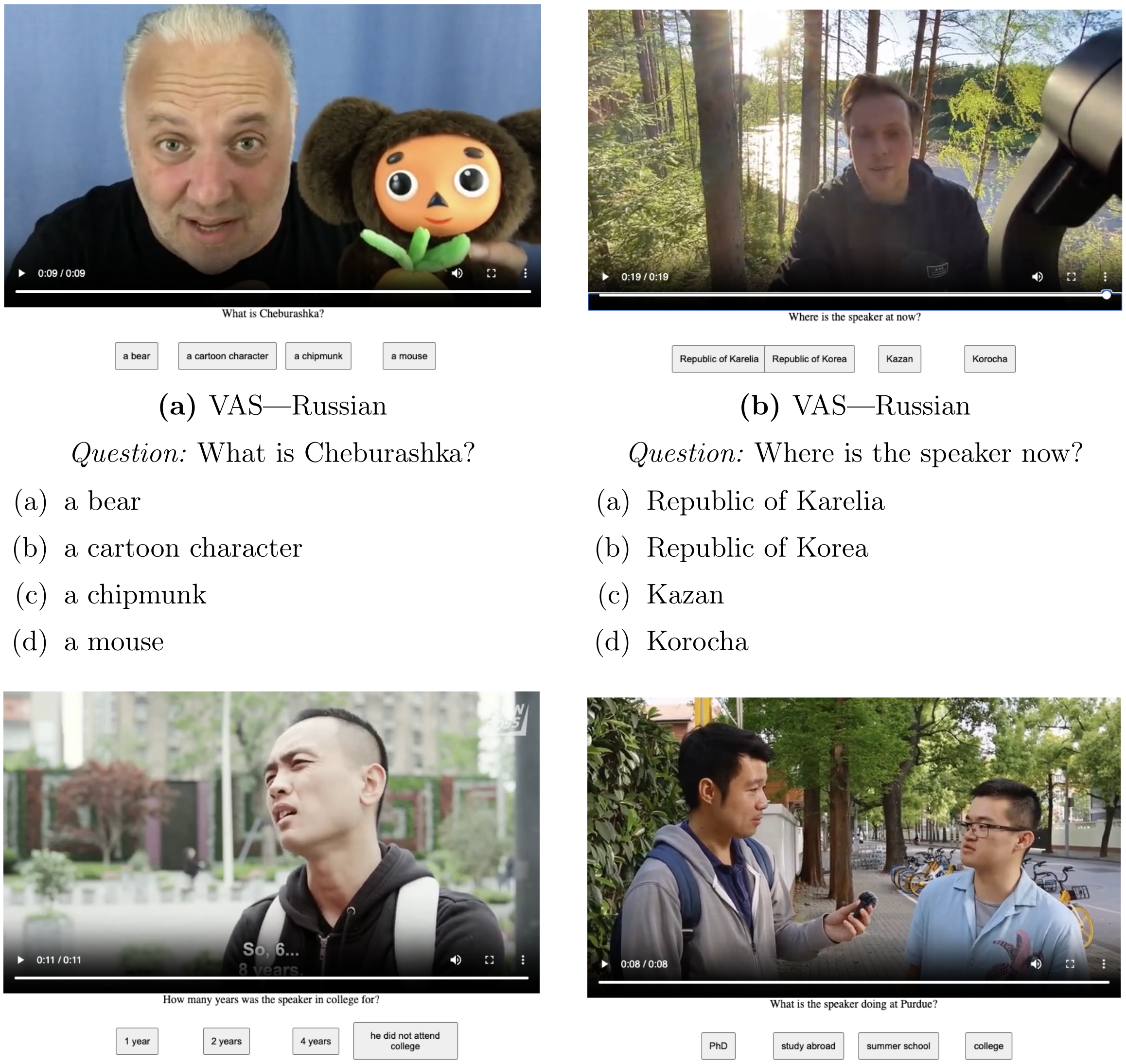
Together, these directions will enhance models of bilingual sociophonetic perception by incorporating richer social data, bridging perception with production, and extending to new populations and methods.

Appendix A

Post-Video Engagement Questions (Examples)

Examples of questions participants saw after watching the social guise video, prompting reflection on speaker identity and background.

Panel A. VAS Task Examples



(a) VAS—Russian

Question: What is Cheburashka?

(a) a bear

(b) a cartoon character

(c) a chipmunk

(d) a mouse

(b) VAS—Russian

Question: Where is the speaker now?

(a) Republic of Karelia

(b) Republic of Korea

(c) Kazan

(d) Korocha

(c) VAS—Chinese

Question: How many years was the speaker in college for?

(a) 1 year

(b) 2 years

(c) 4 years

(d) he did not attend college

(d) VAS—Chinese

Question: What is the speaker doing at Purdue?

(a) PhD

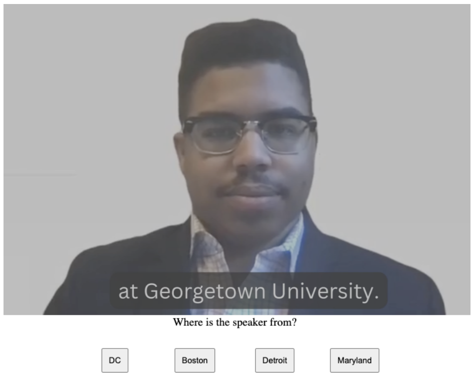
(b) study abroad

(c) summer school

(d) college

Figure A.1: Examples of post-video engagement questions in the VAS task. Four representative screenshots (two Russian, two Chinese) with the exact question prompts and multiple-choice answers shown to participants.

Panel B. AEM Task Examples


at Georgetown University.

Where is the speaker from?

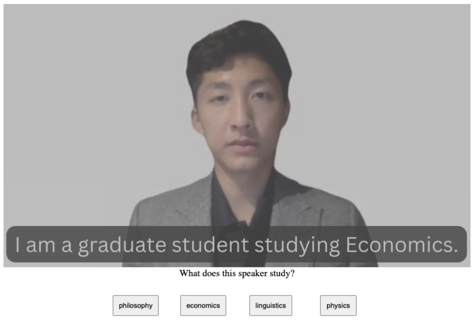
AEM—American Guise

Question: Where is the speaker from?

(a) DC

(b) Boston

(c) Detroit


I am a graduate student studying Economics.

What does this speaker study?

AEM—Chinese Guise

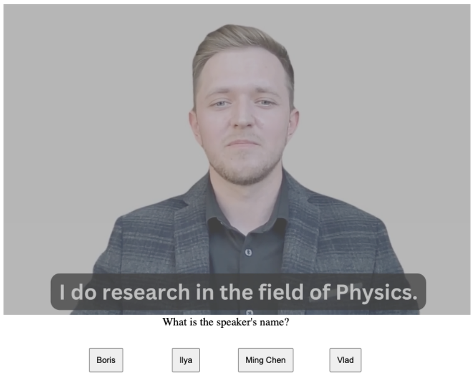
Question: What does the speaker study?

(a) philosophy

(b) economics

(c) linguistics

(d) physics


I do research in the field of Physics.

What is the speaker's name?

AEM—Russian Guise

Question: What is the speaker's name?

(a) Boris

(b) Ilya

(c) Ming Chen

(d) Vlad

Figure A.2: Examples of post-video engagement questions in the AEM task. Each row shows a screenshot from the AEM condition video and the corresponding multiple-choice question displayed to participants.

Appendix B

Individual Rating Patterns Across Chinese and Russian Listeners

This appendix presents individual-level rating trajectories from the Visual Analogue Scale (VAS) task for all participants who completed both baseline and social guise conditions. Each subfigure shows how a participant's voicing categorization changed across the nine VOT steps under three guise conditions (American, Mandarin, Russian). The smoothed trend lines (*loess*) visualize how each listener's responses varied across the continuum, revealing both individual differences and general cross-group patterns.

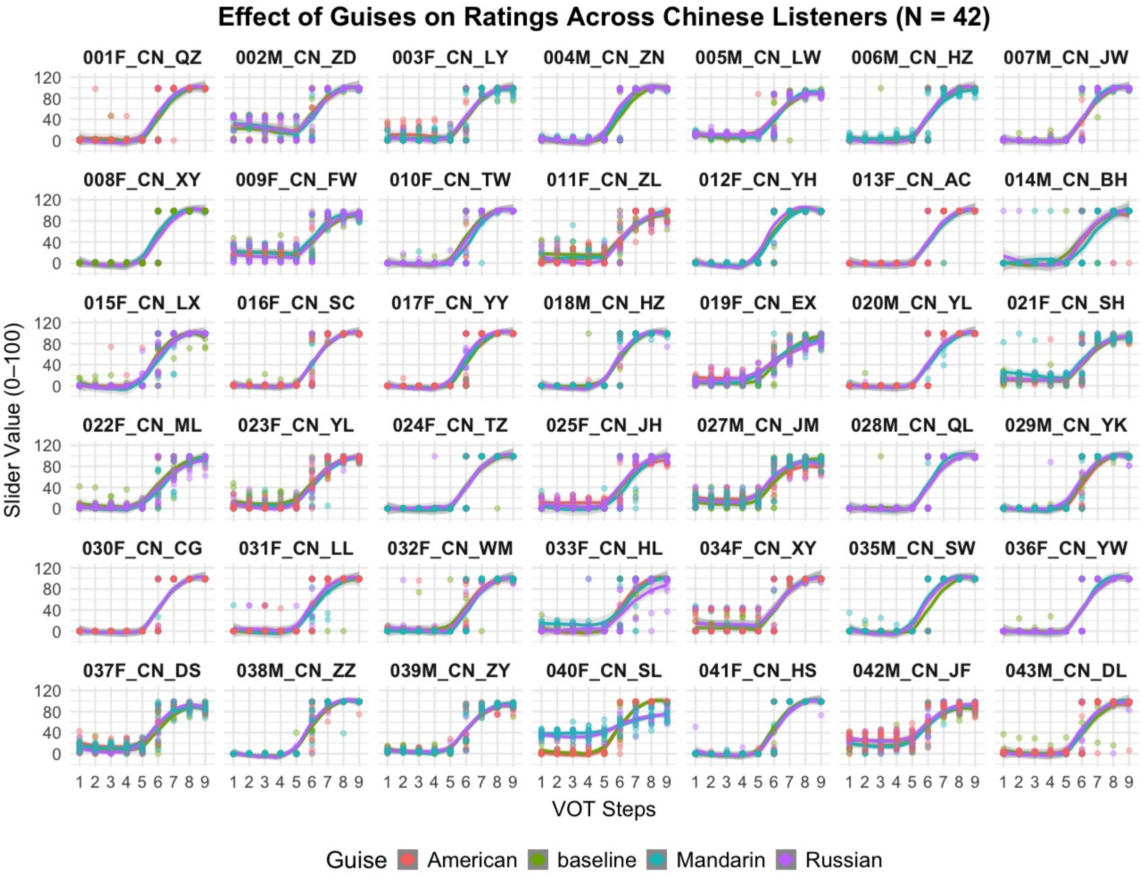


Figure B.1: Panel A. Chinese listeners' individual rating trajectories. Each panel represents one listener's slider responses across the nine VOT steps for all guises. LOESS trend lines with 95% confidence bands illustrate how categorization patterns shift across the continuum within and across guises.

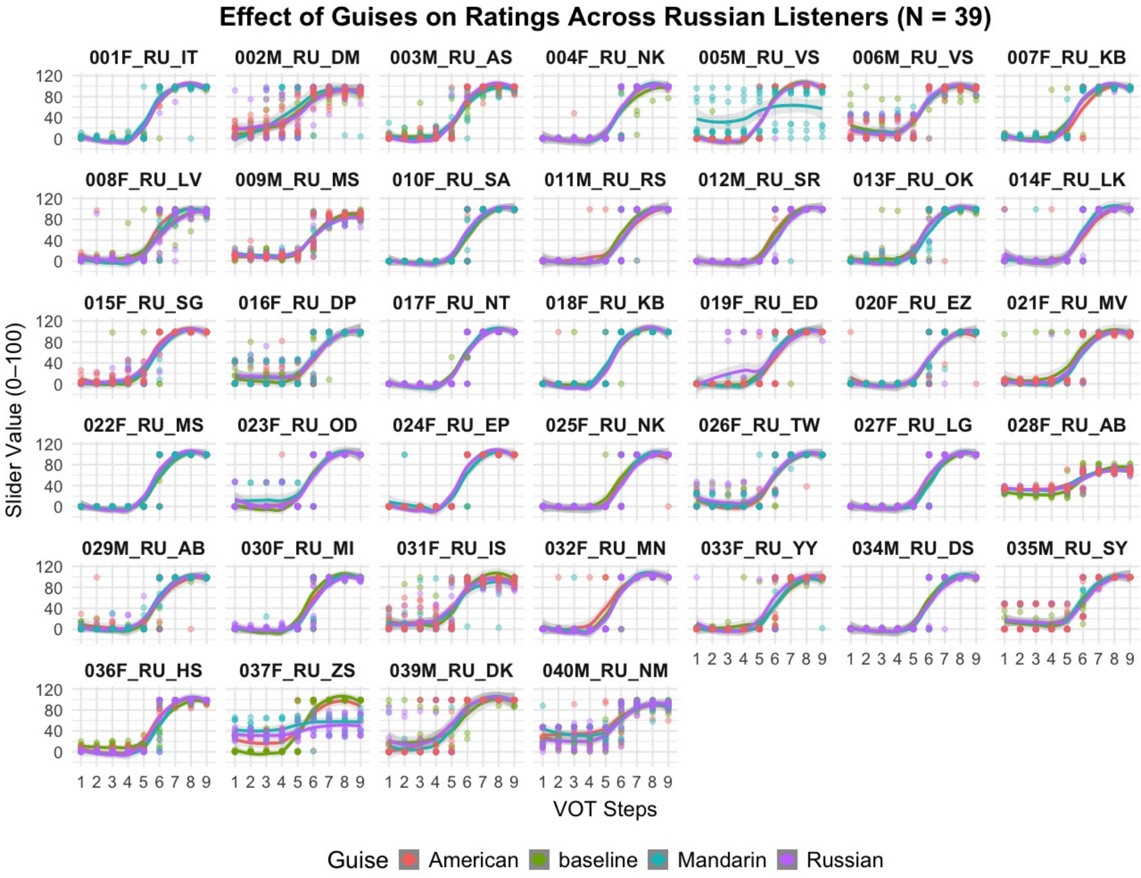


Figure B.2: Panel B. Russian listeners' individual rating trajectories. Each panel shows one listener's VAS ratings across the nine VOT steps for all guises. The individual trends reveal that, while most participants exhibit S-shaped categorization curves, some listeners (e.g., `028F_RU_AB`) display flatter or atypical response patterns, suggesting that their ratings clustered around the center despite clear clarity.

Appendix C

Target Images Used in AEM Experimental Trials

All six pictures used as visual targets for lexical items: **BARK**, **PARK**, **DART**, **TART**, **GUARD**, **CARD**. Images were matched for concreteness, color balance, and recognizability.

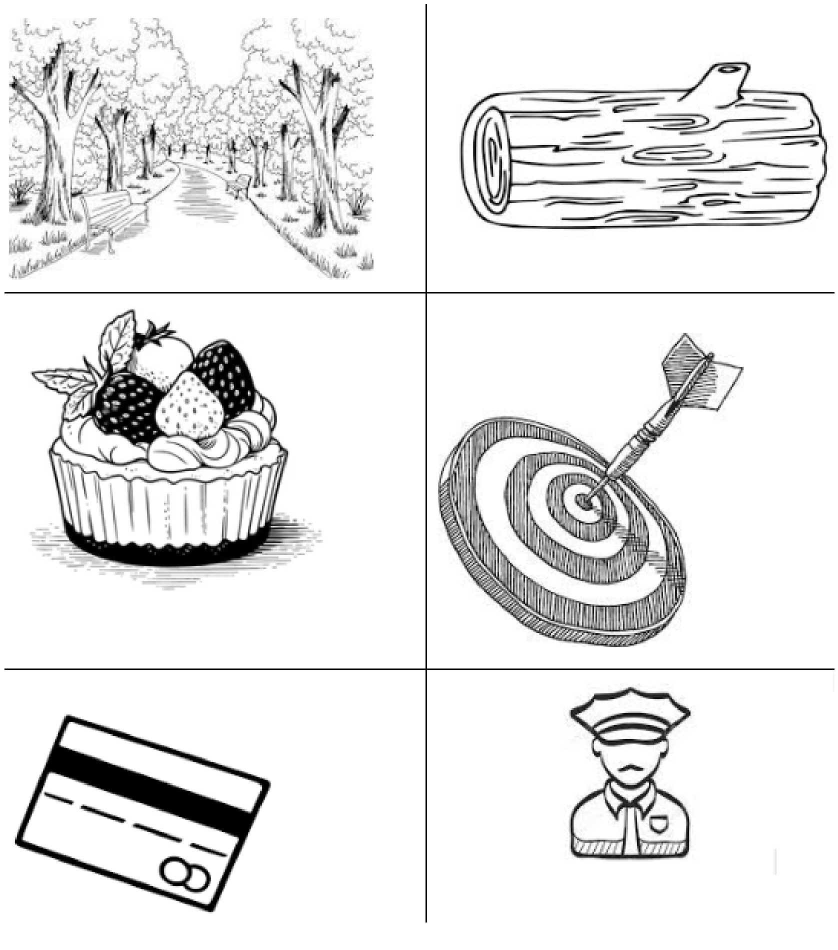
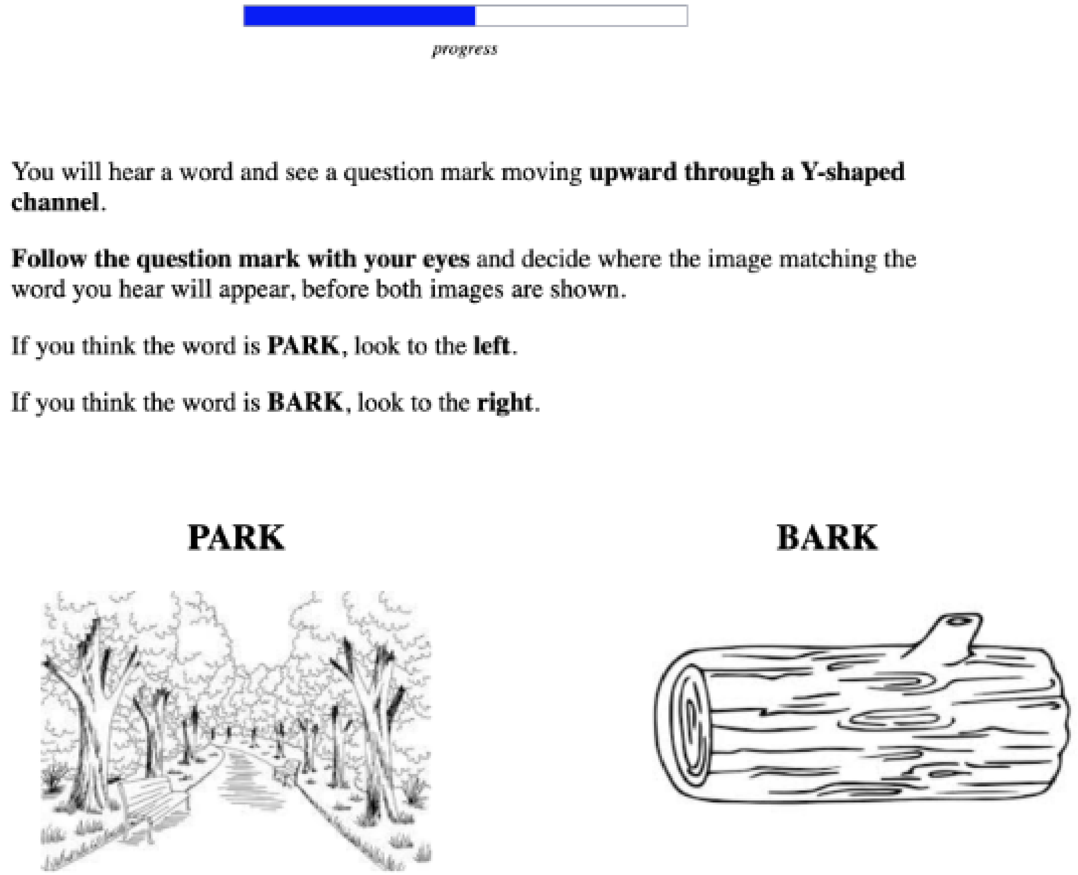


Figure C.1: Target images used in experimental trials. All six visual targets corresponding to the lexical items are displayed together in their matched pairs: *BARK–PARK*, *DART–TART*, and *GUARD–CARD*.

Appendix D

Example Instruction Screen from AEM Task (BARK–PARK phase)



You will hear a word and see a question mark moving **upward through a Y-shaped channel**.

Follow the question mark with your eyes and decide where the image matching the word you hear will appear, before both images are shown.

If you think the word is **PARK**, look to the **left**.

If you think the word is **BARK**, look to the **right**.

PARK

BARK

Figure D.1: Example instruction screen from AEM (*BARK–PARK* Phase). A screenshot of the task introduction presented to participants prior to the *BARK–PARK* trial block. This screen familiarized listeners with the experimental interface and provided instructions on how to respond to auditory stimuli during each trial.

Appendix E

Method Validation: Screen Gaze Location Recovery Test

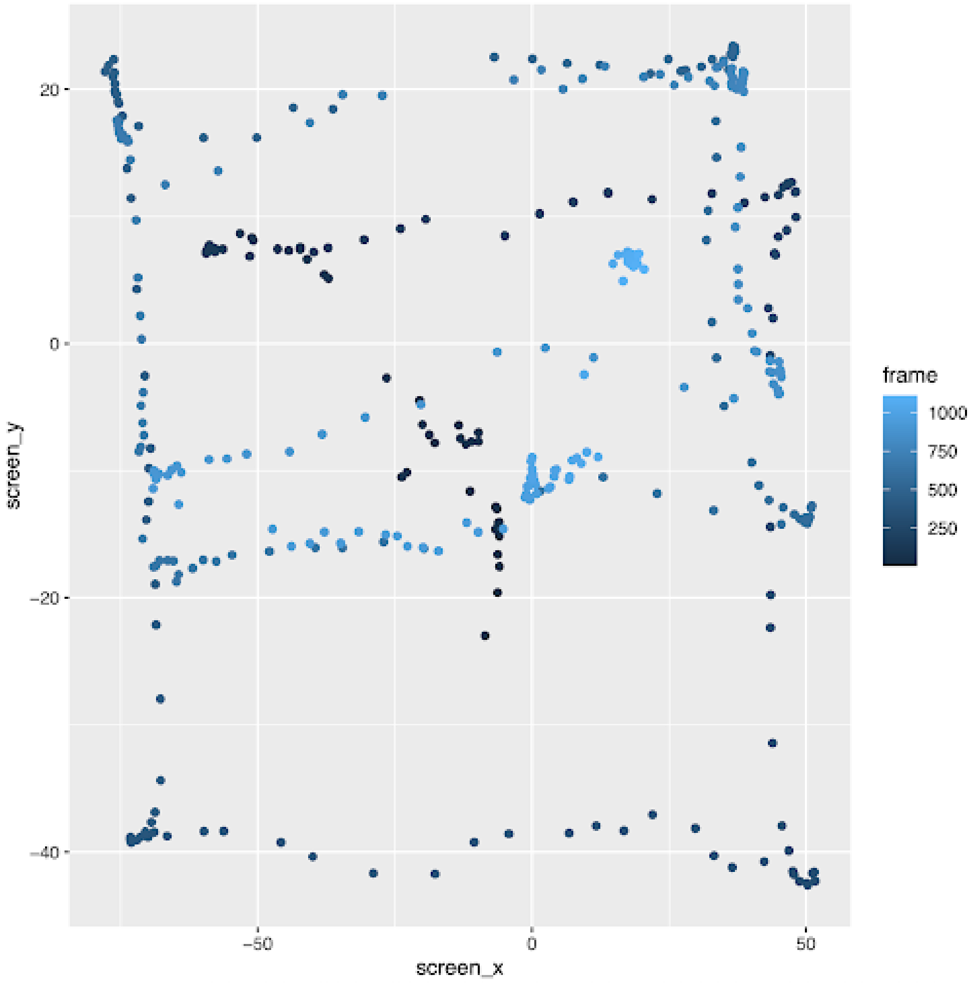


Figure E.1: Recovered screen gaze locations during a test trial. A participant traced a rectangle on the screen using eye movements while also moving the head to make the scenario more realistic and challenging. This qualitative test shows the recovered gaze trajectory and demonstrates the robustness of the gaze estimation method to natural head pose variation.

Appendix F

Voiceless Categorization by Condition (Supplemental Plot)

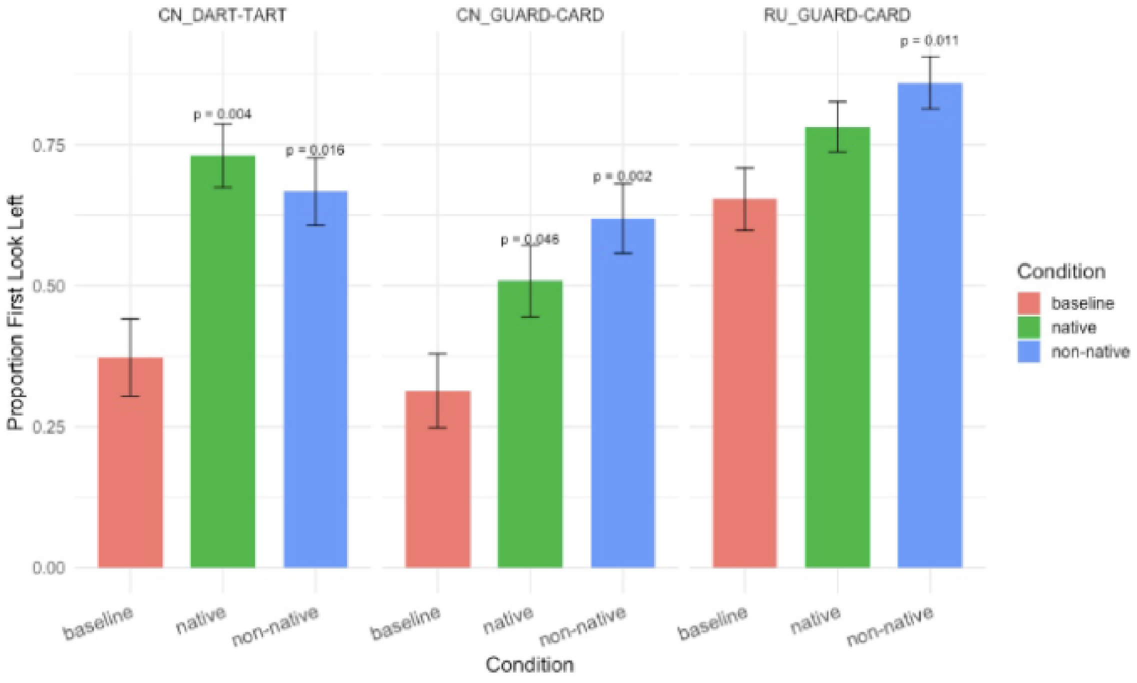


Figure F.1: Proportion of first leftward looks (voiceless categorization) by condition for all significant group × PoA pairs, at ambiguous VOT step 6. P-values from the ambiguous step model are labeled above the relevant bars.

Appendix G

Descriptive Statistics for First Look Categorization in the AEM TASK

Table G.1: Mean and standard deviation of voiceless categorization by condition (Significant groups only).

Group × POA	Condition	Mean Proportion	SD
CN_DART-TART	baseline	0.373	0.488
CN_DART-TART	native	0.730	0.447
CN_DART-TART	non-native	0.667	0.475
CN_GUARD-CARD	baseline	0.314	0.469
CN_GUARD-CARD	native	0.508	0.504
CN_GUARD-CARD	non-native	0.619	0.490
RU_GUARD-CARD	baseline	0.653	0.479
RU_GUARD-CARD	native	0.782	0.416
RU_GUARD-CARD	non-native	0.860	0.350

Appendix H

Mixed-Effects Modeling Results for All Listeners (EIT × AEM, Four-Way Model)

Table H.1: Significant fixed effects for all listeners (Four-way linear mixed effects model).

Term	Estimate	Std. Error	t-value	p-value
(Intercept)	-48.22	21.53	-2.24	.026
vot_step	13.76	3.06	4.50	< .001
exp_conditionAEM-native	119.70	23.54	5.08	< .001
exp_conditionAEM-non-native	62.94	23.55	2.67	.008
listener_groupRussian	61.82	25.34	2.44	.015
vot_step : exp_conditionAEM-native	-18.62	4.17	-4.46	< .001
vot_step : exp_conditionAEM-non-native	-9.53	4.17	-2.28	.022

Continued on next page

Table H.1 (continued)

Term	**Estimate**	**Std. Error**	**t-value**	**p-value**

`EIT :`
`exp_conditionAEM-native`
	-1.14	0.25	-4.59	$<$ .001

`EIT :`
`exp_conditionAEM-non-native`
	-0.50	0.25	-2.01	.045

`vot_step :`
`listener_groupRussian`
	-8.87	3.58	-2.48	.013

`EIT :`
`listener_groupRussian`
	-0.62	0.26	-2.36	.019

`exp_conditionAEM-native :`
`listener_groupRussian`
	-146.0	27.76	-5.26	$<$ .001

`exp_conditionAEM-non-native :`
`listener_groupRussian`
	-103.2	28.39	-3.64	$<$ .001

`vot_step :`
`EIT :`
`exp_conditionAEM-native`
	0.18	0.04	4.03	$<$ .001

`vot_step :`
`EIT :`
`listener_groupRussian`
	0.08	0.04	2.27	.023

Continued on next page

Table H.1 (continued)

Term	**Estimate**	**Std. Error**	**t-value**	**p-value**

`vot_step :`
`exp_conditionAEM-native :`
`listener_groupRussian`
	24.61	4.91	5.01	$< .001$

`vot_step :`
`exp_conditionAEM-non-native :`
`listener_groupRussian`
	18.65	5.02	3.71	$< .001$

`EIT :`
`exp_conditionAEM-native :`
`listener_groupRussian`
	1.41	0.29	4.91	$< .001$

`EIT :`
`exp_conditionAEM-non-native :`
`listener_groupRussian`
	0.98	0.29	3.34	$< .001$

`vot_step :`
`EIT :`
`exp_conditionAEM-native :`
`listener_groupRussian`
	-0.24	0.05	-4.66	$< .001$

`vot_step :`
`EIT :`
`exp_conditionAEM-non-native :`
`listener_groupRussian`
	-0.18	0.05	-3.39	$< .001$

Appendix I

Individual and Group Means Across Experimental Conditions with Participant IDs Labeled

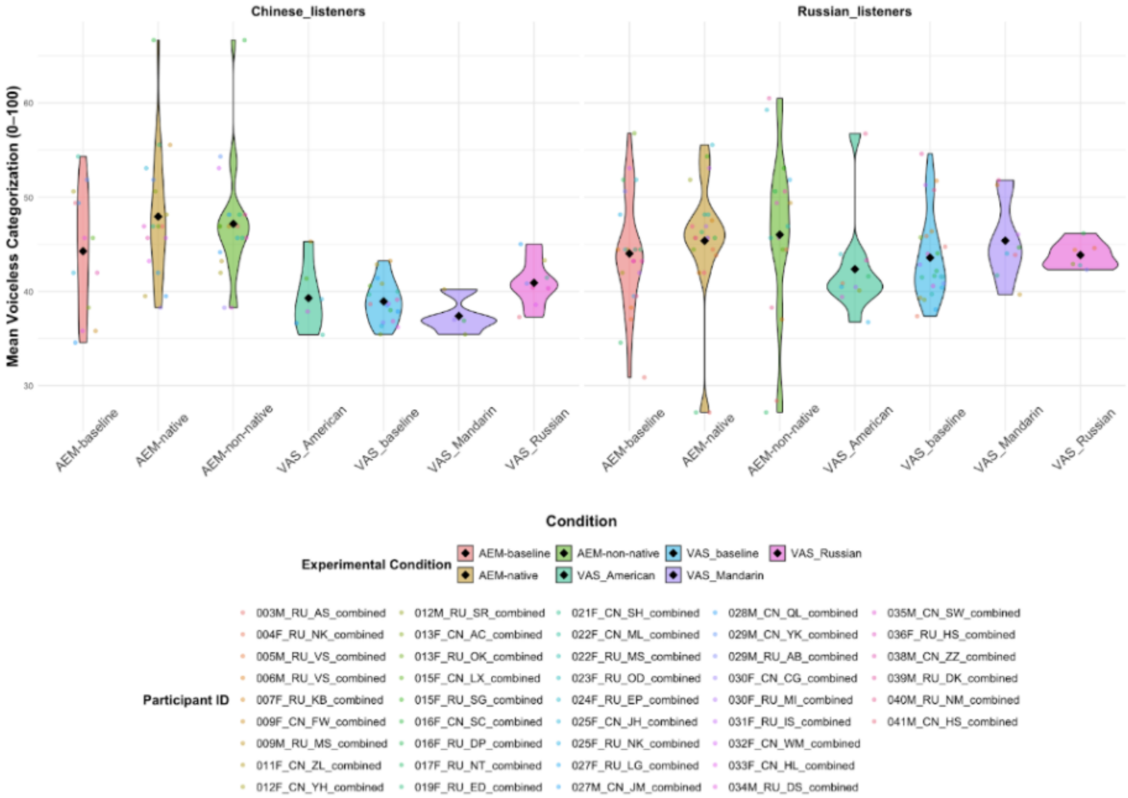


Figure I.1: Individual and group means across experimental conditions with participant IDs labeled. This figure visualizes mean voiceless categorization (0–100 scale) across all seven experimental and guise conditions (`AEM-baseline`, `AEM-native`, `AEM-non-native`, `VAS-baseline`, `VAS-American`, `VAS-Mandarin`, `VAS-Russian`) for both Chinese and Russian listener groups. Unlike the Chapter 5 version, each participant ID is explicitly displayed, allowing individual trajectories to be visually tracked across tasks and guises. This design highlights within-participant consistency and between-group variability in responses while enabling verification that each participant completed all conditions.

Appendix J

Distribution of Numeric AQ Subscale Scores by Listener

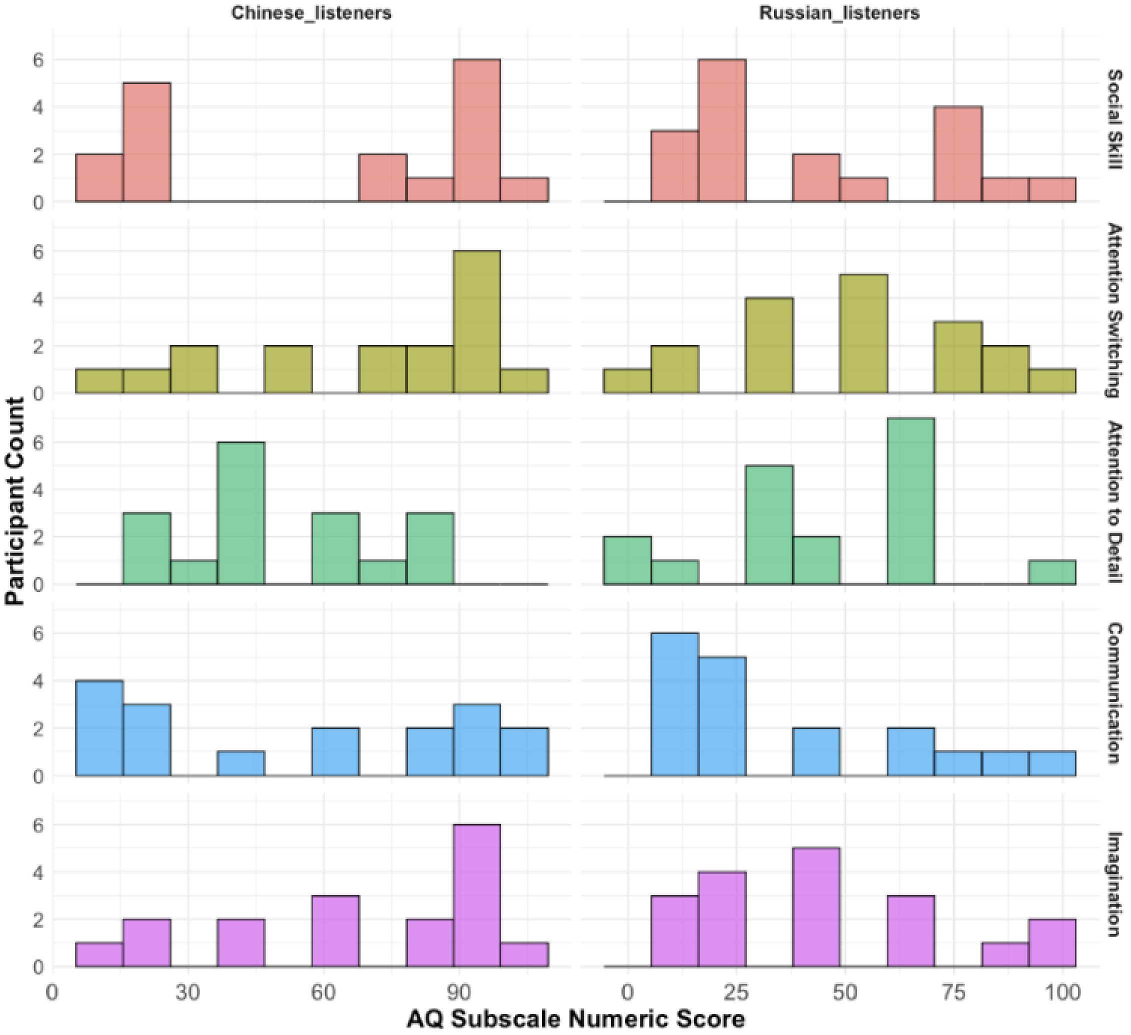
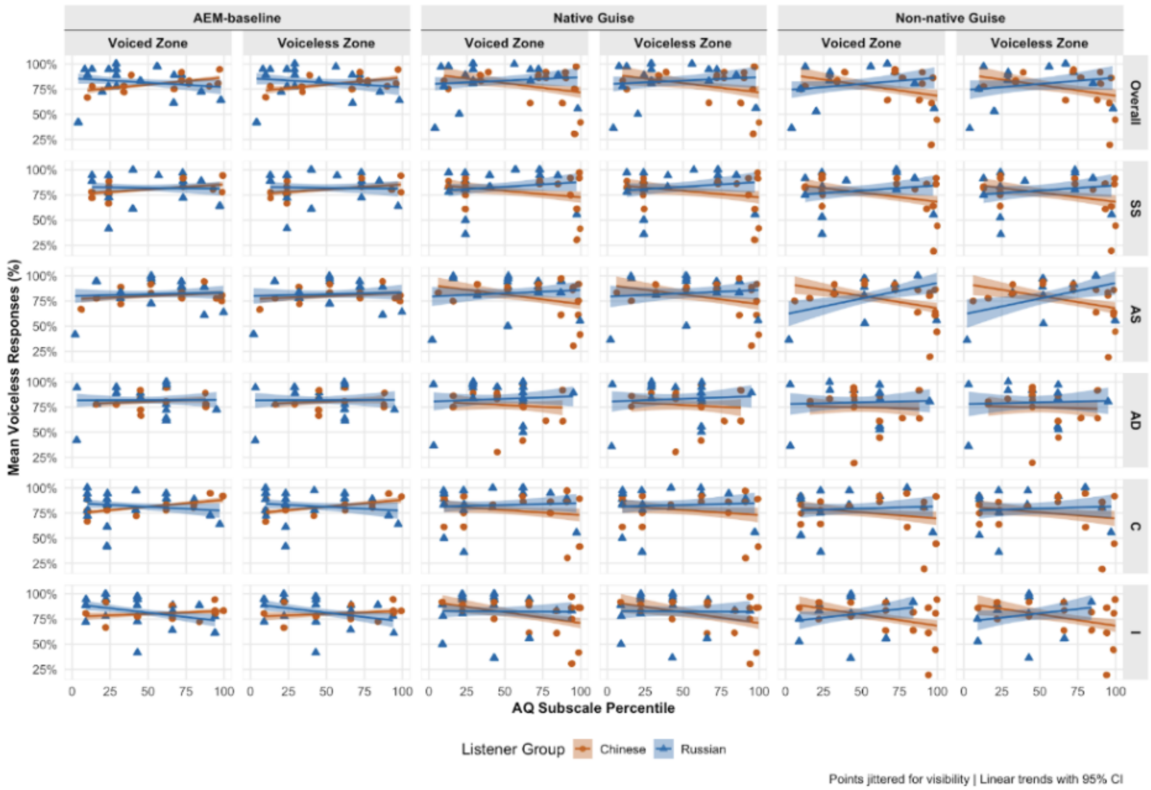


Figure J.1: Distribution of numeric AQ subscale scores by listener group. Histograms show the distribution of scores across five AQ subscales — *social skills, attention switching, attention to detail, communication, and imagination* — for Chinese (left) and Russian (right) listeners.

Appendix K

Mean Voiceless Categorization by AQ Subscale, Condition, and Zone



Points jittered for visibility | Linear trends with 95% CI

Figure K.1: Mean voiceless categorization by AQ subscale, condition, and zone. Plots show mean voiceless responses (%) as a function of AQ subscale percentile across six dimensions (`Overall, Social Skills (SS), Attention Switching (AS), Attention to Detail (AD), Communication (C), Imagination (I)`) and three experimental conditions (`AEM-baseline, AEM-native guise, AEM-non-native guise`), separated into voiced and voiceless zones. Chinese listeners (orange) and Russian listeners (blue) are displayed with linear trends and 95% confidence intervals, illustrating how subscale traits relate to voicing categorization patterns across conditions.

Bibliography

Jean E. Andruski, Sheila E. Blumstein, and Martha Burton. The effect of subphonetic differences on lexical access. *Cognition*, 52(3):163–187, 1994. doi: 10.1016/0010-0277(94)90042-6.

Molly Babel. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177–189, 2012. ISSN 0095-4470. doi: 10.1016/j.wocn.2011.09.001.

Simon Baker, Ben Buchanan, David Hegarty, Carla Smyth, and Emerson Bartholomew. A review of the clinical utility and psychometric properties of the autism spectrum quotient children's version (AQ-child): Gender-specific norms, percentile rankings, and qualitative descriptors. 2025.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17, 2001. ISSN 1573-3432. doi: 10.1023/A:1005653411471.

Catherine T. Best. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, 167(224):233–277, 1994.

Ocke-Schwen Bohn and James Emil Flege. Perceptual switching in spanish/english bilinguals. *Journal of Phonetics*, 21(3):267–290, 1993. ISSN 0095-4470. doi: 10.1016/S0095-4470(19)31339-7.

Harriet Wood Bowden. Assessing second-language oral proficiency for research: The spanish elicited imitation task. *Studies in second language acquisition*, 38 (4):647–675, 2016. Publisher: Cambridge University Press.

Kathryn Campbell-Kibler. Accent,(ING), and the social logic of listener perceptions. *American speech*, 82(1):32–64, 2007. Publisher: Duke University Press.

Alfonso Caramazza, Grace Yeni-Komshian, Edgar Zurif, and Ettore Carbone. Perception and production of stops in bilinguals and unilinguals. *The Journal of the Acoustical Society of America*, 53(1):369–369, 1973. Publisher: Acoustical Society of America.

Laura Staum Casasanto. Does social information influence sentence processing? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30, 2008. Issue: 30.

Dorien Ceuleers, Ingeborg Dhooge, Sofie Degeest, Hanneleen Van Steen, Hannah Keppler, and Nele Baudonck. The effects of age, gender and test stimuli on visual speech perception: a preliminary study. *Folia Phoniatrica et Logopaedica*, 74(2): 131–140, 2022.

Kuan-Yi Chao and Li-mei Chen. A cross-linguistic study of voice onset time in stop consonant productions. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 13, Number 2, June 2008*, pages 215–232, 2008.

Taehong Cho and Peter Ladefoged. Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics*, 27(2):207–229, 1999. Publisher: Elsevier.

Meghan Clayards, Michael K. Tanenhaus, Richard N. Aslin, and Robert A. Jacobs. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809, 2008. Publisher: Elsevier.

Matthew H. Davis and Ingrid S. Johnsrude. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research*, 229(1):132–147, 2007. Publisher: Elsevier.

Xinran Dong. Requests in academic settings in english, russian and chinese, 2009.

Annette D'Onofrio. Controlled and automatic perceptions of a sociolinguistic marker. *Language Variation and Change*, 30(2):261–285, 2018. Publisher: Cambridge University Press.

Penelope Eckert. *Meaning and linguistic variation: The third wave in sociolinguistics.* Cambridge University Press, 2018.

Jeffrey L. Elman, Randy L. Diehl, and Susan E. Buchwald. Perceptual switching in bilinguals. *The Journal of the acoustical Society of America*, 62(4):971–974, 1977. Publisher: Acoustical Society of America.

Robert J. Fisher. Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 20(2):303–315, 1993. Publisher: The University of Chicago Press.

James E. Flege. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92 (1):233–277, 1995. Publisher: York Press.

James Emil Flege, Murray J. Munro, and Ian RA MacKay. Effects of age of second-language learning on the production of english consonants. *Speech Communication*, 16(1):1–26, 1995. Publisher: Elsevier.

Nelson Flores and Jonathan Rosa. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard educational review*, 85(2): 149–171, 2015. Publisher: Harvard Education Publishing Group.

Melinda C. Freyaldenhoven, Donna Fisher Smiley, Robert A. Muenchen, and Tiffany N. Konrad. Acceptable noise level: Reliability measures and comparison to preference for background sounds. *Journal of the American Academy of Audiology*, 17(9):640–648, 2006. ISSN 1050-0545. doi: 10.3766/jaaa.17.9.3.

Adrian Garcia-Sierra, Randy L. Diehl, and Craig Champlin. Testing the double phonemic boundary in bilinguals. *Speech communication*, 51(4):369–378, 2009. Publisher: Elsevier.

Susan Shur-Fen Gau. The adult autism spectrum quotient, aq [taiwanese chinese translation], 2024. URL `https://docs.autismresearchcentre.com/tests/AQ_Adult_Chinese_Taiwan.pdf`. Taiwanese Chinese translation of the Adult Autism Spectrum Quotient (AQ). Accessed 2025-10-16.

Howard Giles and Peter Powesland. Accommodation theory. In Nikolas Coupland and Adam Jaworski, editors, *Sociolinguistics*, pages 232–239. Macmillan Education UK, 1997. ISBN 978-0-333-61180-7 978-1-349-25582-5. doi: 10.1007/978-1-349-25582-5_19.

Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68, 1991. Publisher: Cambridge.

Ksenia Gnevsheva. *Variation in passing for a native speaker: Accentedness in second language speakers of English in production and perception.* PhD thesis, University of Canterbury, 2015.

Ksenia Gnevsheva. The expectation mismatch effect in accentedness perception of asian and caucasian non-native speakers of english. *Linguistics*, 56(3):581–598, 2018. ISSN 0024-3949, 1613-396X. doi: 10.1515/ling-2018-0006.

Stephen D. Goldinger. Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem. *Perception & Psychophysics*, 60(6):952–965, 1998. ISSN 0031-5117, 1532-5962. doi: 10.3758/BF03211931.

Julie D. Golomb, Jonathan E. Peelle, and Arthur Wingfield. Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *The Journal of the Acoustical Society of America*, 121(3):1701–1708, 2007. Publisher: AIP Publishing.

Jean K. Gordon, Kim Andersen, Gabriella Perez, and Eileen Finnegan. How old do you think i am? speech-language predictors of perceived age and communicative competence. *Journal of Speech, Language, and Hearing Research*, 62(7):

2455–2472, 2019. ISSN 1092-4388, 1558-9102. doi: 10.1044/2019_JSLHR-L-19-0025.

Jennifer Hay and Katie Drager. Stuffed toys and speech perception. *Linguistics*, 48(4), 2010.

Jennifer Hay, Aaron Nolan, and Katie Drager. From fush to feesh: exemplar priming in speech perception. *Linguistic review*, 23(3), 2006.

Fan Jiang and Shelia Kennison. The impact of l2 english learners' belief about an interlocutor's english proficiency on l2 phonetic accommodation. *Journal of Psycholinguistic Research*, 51(1):217–234, 2022. ISSN 0090-6905, 1573-6555. doi: 10.1007/s10936-021-09835-7.

Keith Johnson, Elizabeth A. Strand, and Mariapaola D'Imperio. Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics*, 27(4):359–384, 1999. Publisher: Elsevier.

Dave F. Kleinschmidt and T. Florian Jaeger. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148, 2015. Publisher: American Psychological Association.

Eun Jong Kong and Jan Edwards. Individual differences in speech perception: Evidence from visual analogue scaling and eye-tracking. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, 2011.

Eunjong Kong and Jan Edwards. Individual differences in l2 learners' perceptual cue weighting patterns. In *ICPhS*, 2015.

Christian Koops, Elizabeth Gentry, and Andrew Pantos. The effect of perceived speaker age on the perception of PIN and PEN vowels in houston, texas. *U.Penn Working Papers in Linguistics*, 14.2, 2008.

Maria Kostromitina and Luke Plonsky. Elicited imitation tasks as a measure of l2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 44(3): 886–911, 2022. Publisher: Cambridge University Press.

Ethan Kutlu. Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different english varieties. *Journal of Multilingual and Multicultural Development*, 44(6):511–525, 2023. ISSN 0143-4632, 1747-7557. doi: 10.1080/01434632.2020.1835929.

Ethan Kutlu, Samantha Chiu, and Bob McMurray. Moving away from deficiency models: Gradiency in bilingual speech categorization. *Frontiers in psychology*, 13:1033825, 2022a. Publisher: Frontiers Media SA.

Ethan Kutlu, Mehrgol Tiv, Stefanie Wulff, and Debra Titone. Does race impact speech perception? an account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications*, 7(1):7, 2022b. ISSN 2365-7464. doi: 10.1186/s41235-022-00354-0.

Ethan Kutlu, Mehrgol Tiv, Stefanie Wulff, and Debra Titone. The impact of race on speech perception and accentedness judgements in racially diverse and non-diverse groups. *Applied Linguistics*, 43(5):867–890, 2022c. Publisher: Oxford University Press.

William Labov. Some principles of linguistic methodology. *Language in Society*, 1(1):97–120, 1972. Publisher: Cambridge University Press.

Wallace E. Lambert, Richard C. Hodgson, Robert C. Gardner, and Samuel Fillenbaum. Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60(1):44, 1960. Publisher: American Psychological Association.

Geoffrey Leech and Tatiana Larina. Politeness: West and east. *Vesnik RUDN, Seriya Lingvistika [Russian Journal of Linguistics]*, (4):32–47, 2014. ISSN 2313-2299.

Shiri Lev-Ari and Sharon Peperkamp. Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics*, 41(5):320–331, 2013. Publisher: Elsevier.

Rosina Lippi-Green. *English with an accent: Language, ideology and discrimination in the United States.* Routledge, 2012.

Leigh Lisker and Arthur S. Abramson. A cross-language study of voicing in initial stops: Acoustical measurements. *WORD*, 20(3):384–422, 1964. ISSN 0043-7956, 2373-5112. doi: 10.1080/00437956.1964.11659830.

Lars-Olov Lundqvist and Helen Lindner. Is the autism-spectrum quotient a valid measure of traits associated with the autism spectrum? a rasch validation in adults with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 47(7):2080–2091, 2017. ISSN 0162-3257, 1573-3432. doi: 10.1007/s10803-017-3128-y.

Mackenzie Marcinko. *Attitudinal Differences Towards Peruvian Quechua and Spanish Speakers: A Matched Guise Study.* University of Delaware, 2023.

Viorica Marian and Michael Spivey. Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism: Language and Cognition*, 6(2):97–115, 2003. Publisher: Cambridge University Press.

Hazel R. Markus and Shinobu Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253, 1991. ISSN 1939-1471. doi: 10.1037/0033-295X.98.2.224. Place: US Publisher: American Psychological Association.

Kevin B. McGowan. Social expectation improves speech perception in noise. *Language and Speech*, 58(4):502–521, 2015. ISSN 0023-8309, 1756-6053. doi: 10.1177/0023830914565191.

Kevin B. McGowan. Sounding chinese and listening chinese: Awareness and knowledge in the laboratory. *Awareness and Control in Sociolinguistic Research*, pages 25–61, 2016. Publisher: Cambridge University Press Cambridge.

Warda Nejjari, Marinel Gerritsen, Roeland Van Hout, and Brigitte Planken. Refinement of the matched-guise technique for the study of the effect of non-native accents compared to native accents. *Lingua*, 219:90–105, 2019. Publisher: Elsevier.

Nancy Niedzielski. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1):62–85, 1999. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X99018001005.

Kuniko Nielsen. Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(2):132–142, 2011. Publisher: Elsevier.

Kuniko Y. Nielsen and Rebecca Scarborough. Perceptual asymmetry between greater and lesser vowel nasality and VOT. In *ICPhS*, 2015.

NovoPsych. Autism spectrum quotient (aq), 2025. URL `https://novopsych.com/assessments/diagnosis/autism-spectrum-quotient/`. Accessed via NovoPsych online assessment library.

John J. Ohala. Phonological evidence for top-down processing in speech perception. In *Invariance and variability in speech processes*, pages 386–401. Psychology Press, 2014.

Jinghua Ou and Sam-Po Law. Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Attention, Perception, & Psychophysics*, 79(3):945–963, 2017. ISSN 1943-3921, 1943-393X. doi: 10.3758/s13414-017-1283-z.

Jinghua Ou, Sam-Po Law, and Roxana Fung. Relationship between individual differences in speech processing and cognitive functions. *Psychonomic Bulletin & Review*, 22(6):1725–1732, 2015. ISSN 1069-9384, 1531-5320. doi: 10.3758/s13423-015-0839-y.

Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *Visual Cognition*, 62(1):145–168, 2009. ISSN 1747-0218, 1747-0226. doi: 10.1080/01690960902825973.

Eva Reinisch and Lori L. Holt. Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2):539, 2014. Publisher: American Psychological Association.

Catherine Ringen and Vladimir Kulikov. Voicing in russian stops: Cross-linguistic implications. *Journal of Slavic linguistics*, 20(2):269–286, 2012. Publisher: Slavica Publishers.

Casey L. Roark, Erica Lescht, Amanda Hampton Wray, and Bharath Chandrasekaran. Auditory and visual category learning in children and adults. *Developmental Psychology*, 59(5):963, 2023. Publisher: American Psychological Association.

Leah Roberts. Individual differences in second language sentence processing. *Language Learning*, 62:172–188, 2012. ISSN 0023-8333, 1467-9922. doi: 10.1111/j.1467-9922.2012.00711.x.

William S. Robinson. Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, 38(2):337–341, 2009. Publisher: Oxford University Press.

Jonathan Rosa and Nelson Flores. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647, 2017. Publisher: Cambridge University Press.

Donald L. Rubin. Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, 33(4):511–531, 1992. ISSN 0361-0365, 1573-188X. doi: 10.1007/BF00973770.

A. P. Shabalin. Shkala autisticheskogo spektra (aq), vzroslaia versiia [autism spectrum quotient (aq), adult version]. Translation, Autism Research Centre, 2024. URL `https://docs.autismresearchcentre.com/tests/AQ_Adult_Russian_v2.pdf`. Russian translation. Accessed 2025-10-16. Available from `https://novopsych.com/assessments/diagnosis/autism-spectrum-quotient/`.

Viktor Shklovsky. Iskusstvo kak priëm. PDF available online, 1917. URL https://www.opojaz.ru/manifests/kakpriem.html. Accessed 10-15-2025.

Viktor Shklovsky. Art as technique. In Lee T. Lemon and Marion J. Reis, editors, *Russian Formalist Criticism: Four Essays*, pages 3–24. University of Nebraska Press, Lincoln, 1965. Original work published 1917.

Mary E. Stewart and Mitsuhiko Ota. Lexical effects on speech perception in individuals with "autistic" traits. *Cognition*, 109(1):157–162, 2008. Publisher: Elsevier.

Elizabeth A. Strand. Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, 18(1):86–100, 1999. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X99018001006.

Meghan Sumner, Seung Kyung Kim, Ed King, and Kevin B. McGowan. The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in psychology*, 4:1015, 2014. Publisher: Frontiers Media SA.

Meredith Tamminga. Matched guise effects can be robust to speech style. *The Journal of the Acoustical Society of America*, 142(1):EL18–EL23, 2017. Publisher: AIP Publishing.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7777863.

Nicole Tracy-Ventura, Kevin McManus, John M. Norris, and Lourdes Ortega. "repeat as much as you can": Elicited imitation as a measure of oral proficiency in l2 french. In Pascale Leclercq, Heather Hilton, and Amanda Edmonds, editors, *Measuring L2 Proficiency: Perspectives from SLA*, pages 143–166. Multilingual Matters, Bristol, UK, 2014.

Jorge R. Valdés Kroff, Rosa E. Guzzardo Tamargo, and Paola E. Dussias. Experimental contributions of eye-tracking to the understanding of comprehension processes while hearing and reading code-switches. *Linguistic Approaches to Bilingualism*, 8(1):98–133, 2018. ISSN 1879-9264, 1879-9272. doi: 10.1075/lab.16011.val.

Charlotte Vaughn and Abby Walker. “i always think that these are funny”: The experience of being a participant in a speaker perception task. In *52nd Annual Meeting of New Ways of Analyzing Variation (NWAV)*, 2024.

Virginia Volterra and Traute Taeschner. The acquisition and development of language by bilingual children. *Journal of Child Language*, 5(2):311–326, 1978. Publisher: Cambridge University Press.

Lacey Wade. Experimental evidence for expectation-driven linguistic convergence. *Language*, 98(1):63–97, 2022. Publisher: Linguistic Society of America.

Lacey Wade, Wei Lai, and Meredith Tamminga. The reliability of individual differences in VOT imitation. *Language and Speech*, 64(3):576–593, 2021. ISSN 0023-8309, 1756-6053. doi: 10.1177/0023830920947769.

Matthew B. Winn. Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script. *The Journal of the Acoustical Society of America*, 147 (2):852–866, 2020. Publisher: AIP Publishing.

Shu-Ling Wu and Lourdes Ortega. Measuring global oral proficiency in SLA research: A new elicited imitation test of l2 chinese. *Foreign Language Annals*, 46(4):680–704, 2013. ISSN 0015-718X, 1944-9720. doi: 10.1111/flan.12063.

Alan CL Yu. Perceptual compensation is correlated with individuals'"autistic" traits: Implications for models of sound change. *PloS one*, 5(8):e11950, 2010. Publisher: Public Library of Science San Francisco, USA.

Alan CL Yu, Julian Grove, Martina Martinovic, and Morgan Sonderegger. Effects of working memory capacity and "autistic" traits on phonotactic effects in speech perception. In *Proceedings of the international congress of the phonetic sciences xvii, Hong Kong: International congress of the phonetic sciences*, pages 2236–2239, 2011.

Magdalena Zając and Arkadiusz Rojczyk. Imitation of english vowel duration upon exposure to native and non-native speech. *Poznan Studies in Contemporary Linguistics*, 50(4):495–514, 2014. Publisher: De Gruyter.

Jérémy Zehr and Florian Schwarz. PennController for internet based experiments (IBEX), 2018. Publisher: OSF.

Adriana A. Zekveld, Dirk J. Heslenfeld, Joost M. Festen, and Ruurd Schoonhoven. Top–down and bottom–up processes in speech comprehension. *Neuroimage*, 32(4):1826–1836, 2006. Publisher: Elsevier.

Georgia Zellou, Delphine Dahan, and David Embick. Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience*, 32(6):776–791, 2017. ISSN 2327-3798, 2327-3801. doi: 10.1080/23273798.2016.1275710.