UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
The Integration of Social and Acoustic Cues During Speech Perception

Permalink
https://escholarship.org/uc/item/5238v03k

Author
Wilbanks, Eric

Publication Date
2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

The Integration of Social and Acoustic Cues During Speech Perception

by

Eric Wilbanks

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Keith Johnson, Chair
Associate Professor Justin Davidson
Professor Sharon Inkelas
Professor Terry Regier

Fall 2022

The Integration of Social and Acoustic Cues During Speech Perception

Copyright 2022
by
Eric Wilbanks

Abstract

The Integration of Social and Acoustic Cues During Speech Perception

by

Eric Wilbanks

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith Johnson, Chair

How do social characteristics of a speaker influence how listeners process their speech? There is evidence that social characteristics, like a speaker's age, gender, and so forth, can shift how listeners respond to their speech. For example, ambiguous sounding words are recognized quicker and more accurately when matched with pictures of speakers who are likely to say those words, and changing a visual cue about a speaker while keeping the audio constant can change listeners' judgments of what they heard. An unanswered question is whether social information is directly affecting perception, or if it is only affecting later decision-making. Addressing this question will contribute to our understanding of the role of social information during speech perception and will further develop our models of human language processing.

Previous work has demonstrated that social cues can directly shift categorization behavior, but the paradigms used in the majority of this work do not allow conclusions to be drawn about the specific time-course of social and acoustic cue integration. The specific time-course of this process is critical to debates over the nature of linguistic representation (episodic vs. abstract) and processing (feed-forward vs. interactive models). This dissertation investigates how listeners use social cues during speech perception in real-time by measuring the cues' influence on the earliest stages of speech perception. In what follows, I ask whether social cues can induce selective adaptation effects in perception and whether social cues can influence on-line perception as detected via eye-tracking during a compensation for coarticulation task.

For the selective adaptation experiments, I replicate previous work and show less "SH" categorizations when listeners are repeatedly exposed to clear exemplars of /ʃ/. While we do observe evidence for potential subtle influences of speaker gender on the magnitude of the selective adaptive effect, we do not observe evidence for our critical prediction: we do not observe an effect of speaker gender guise on the direction of selective adaptation. That is, the selective behavior of an ambiguous [s]-[ʃ] is not shifted by the perceived gender of a speaker. Additionally, we observe no evidence for the role of

visual face gender cues or acoustic cues to speaker sexuality on the selective adaptation behavior of listeners.

Turning to the compensation for coarticulation eye-tracking experiment, I again replicate the classical effect. For both *asta-ashka* and *alda-arga* stimuli, we observe perceptual compensation for the specific acoustics of C1 in listeners' C2 categorizations. Additionally, this effect emerges gradually as the stimuli unfold, with observable gradient differences between listeners' fixation data developing even before the stimulus has been completed. We do not, however, observe evidence for our core prediction of speaker gender influencing the compensation behavior of the ambiguous C1 step on later C2 categorization.

Taken as a whole, these results indicate that the role of sociophonetic cues in perception may be restricted to later decision stages, rather than exerting their influence during earlier perceptual stages. Though further investigations are required for a more robust conclusion, the experiments detailed in this dissertation present critical evidence for the precise role of these sociophonetic cues in our understanding of speech perception.

To my parents,

who gave me
the best
parts of themselves.

Contents

List of Figures

List of Tables

Acknowledgments

Well team, we made it!

First off, I want to thank my committee who have been phenomenal sources of support not only during this dissertation, but also throughout my time at Berkeley.

Keith Johnson has been an absolute font of knowledge and wisdom, of course. There hasn't been a question or problem I've faced that wasn't addressed by a quick coffee chat with Keith. But during my time at Berkeley, this shared knowledge and expertise was the characteristic of Keith that I valued second. What I've valued first of all is how kind and openhearted Keith has been, and his willingness to provide emotional support and listen to my worries and anxieties helped get me through this stressful Ph.D. journey. Before I arrived at Berkeley, someone I really respect told me about a time where Keith showed them a simple but deeply meaningful moment of kindness. This anecdote helped convince me that Berkeley was a department where I would be welcome. Thanks Keith for being your kind and supportive self, in the small moments and the big ones.

Coming directly into this program from a department focused on sociolinguistics, Justin Davidson helped me feel right at home. His supportive and expert advice on all things socio significantly supported my continued training and informed my approach to the research in this dissertation. I also really appreciate the welcoming and collaborative socio research group Justin developed and continues to nurture! Although I ended up doing less research on Spanish than I initially thought, Justin has continued to be there to support and guide me along. ¡Muchísimas gracias, te lo agredezco mucho!

Another welcoming and supportive presence in my training has been Sharon Inkelas. Sharon helped expand my knowledge of formal phonology, and embodied the spirit of breadth of inquiry that Berkeley Linguistics champions. In addition, she was an incredibly supportive mentor throughout my time at Berkeley, even encouraging me to collaborate on a research project in formal phonology, an incredibly rewarding and fun experience for me. When I consider some of my most lasting memories of Sharon though, my mind turns to her practical advice on balancing being a scholar and an instructor in LING375. Sharon cheered for, encouraged, and informed our small cohort of new GSIs and was exactly the voice we needed to make it through a trying new experience.

This dissertation and my training at Berkeley was also significantly shaped by the guidance and influence of Terry Regier. I've had the immense pleasure of working with Terry in many different capacities: through several incredibly engaging and impactful research seminars, as a Graduate Student Instructor, and on service committees. Throughout all of these varied contexts, Terry has remained a consistent exemplar of the power of thoughtful and careful consideration. His practical strategy of defining a minimal viable product, a slightly more ambitious goal, and a pie-in-the sky best case scenario is a practice which I continue to model in goal-setting to this day. This thoughtful and practical approach continued to his mentoring of this dissertation research. Whenever I would get too focused in on one particular methodological hitch, or a narrow and overly specific aspect of some framing text, Terry could always be counted on to ask for a moment of pause, and then guide me back to considering the core question and impact of the work.

This dissertation was also made possible by the support of other amazing faculty and staff at Berkeley. Susan Lin was such a thoughtful and supportive advisor and mentor for my first few years of the program and her Berkeley pitch was a large part of me finally deciding to attend. Throughout our time as colleagues, Susan constantly fostered an atmosphere of engaged and supportive scholarship that really helped me develop as a researcher, instructor, and person. I hope I can recreate that in the future for others!

If you ask anyone in the department, they'll tell you that Ron Sprouse is a wizard. Luckily for us, Ron is less a reclusive-hermit-in-the-woods type wizard and more of a Gandalf-rides-to-Helm's-Deep type wizard (but for all our linguistics tech problems). Whether it was troubleshooting a tricky bit of analysis or getting a last minute eyetracking experiment off the ground, the work in this dissertation (and the department generally) literally could not be done without Ron's knowledge and patience.

I'd like to also thank the amazing Belén Flores, Paula Floro, and Martine Alexander who spearheaded the essential functions of the department. In addition to their encyclopedic knowledge of every aspect of doing research and teaching in the University, Belén, Paula, and Martine also blessed us with their incredible levels of care and attention to our overall well-being and status as people, not just researchers. I always knew that they would have the answers to any questions I had, as well as to questions I didn't even think to have. When Belén and Paula moved off to the next phases of their retired life (congratulations!), Johnny Morales Arellano and Siti Keo eagerly stepped in to help us continue on, and I'd like to thank them both for being there in our time of need!

I'd also like to thank those that assisted in the logistics of getting this project off the ground. Thank you to XLab and especially Rowilma Balza del Castillo for the final participant recruiting push. I'd also like to thank those who provided their voices as the starting point for the stimuli in this dissertation (you know who you are!). This project and others were supported by some of the most inspiring undergraduate LRAP mentees and I am so grateful for having had the chance to work with you all over the years; a special thank you to Cecilia Gao and Andres Sanchez, who directly assisted with the piloting of the eyetracking equipment used in this dissertation, you all are awesome! An enormous and sincere thank you to Agneta Herlitz and Martin Asperholm who graciously shared their face stimuli, allowing for an exciting experiment that would have been infeasible for me on my own. Finally, a giant shout out to my Summer 2022 Writing Group who provided valuable feedback (and commiseration) during the drafting process: Carolina Talavera, David I. Delano, and Levi Vonk.

Throughout my journey to this point, I was shaped and supported by some amazing colleagues and mentors. First off, thank you to my friends and mentors at NCSU Linguistics who first introduced me to linguistics and provided such a vibrant environment to learn, research, and grow. Special shout outs to Robin Dodsworth, Jeff Mielke, Erik Thomas, and Walt Wolfram for being such inspiring mentors! Thank you to my colleagues and mentors at Berkeley Linguistics, who helped me broaden my horizons and further develop myself. In addition to those I mentioned above, I want to thank the Sociolinguistics Lab at Berkeley for being a welcoming home for discussion and fun, as well as all of the members of the PhonLab for always being down to look at an interesting spectrogram, complain about an experiment that wasn't working, and making our office feel cozy and welcome (despite the lack of windows).

Thanks to my amazing cohort: Karee, Myriam, Noah, Tessa, and Yevgeniy. It's been a *journey* since we all started and I'm happy I got to do it with you all! (Though I could have done without the stress of BLS, but I guess trial-by-fire is still a useful bonding activity...). I'm looking forward to our next cohort tea, once we're all back together.

Of course, getting to this point would not have been possible without the support, love, and patience of my family and friends. Shout out to my D&D crews who have been a source of laughter, fun, and hijinks over the years, and especially at the peak of pandemic lockdown. I'd also like to thank my Raleigh friends, Lindsey, David, Colleen, and Tyler, for continuing to be positive forces in my life, even from so far away. My time at Berkeley has also been enriched by the friendship of Alice, Raksit, Edwin, and Zach who were always down for boba, great food, and complaining about work. None of you are allowed to move, sorry!

Thank you to all of my extended family for their support and love, and for listening when I explain what exactly it is that I've been doing over here for so long. Even though we're far apart, catching up with my siblings whenever I could was such an immense support for me during this process. Special shout out to Justin and Paxton for letting me complain about work and play with their adorable pets when I visited.

Thank you to my Dad and Stepmom, who are constant anchors in my life and sources of support and inspiration. I know that I could always count on them to cheer me on, give great advice, and always be ready for me to come home to visit.

Thank you to my Mom, for being a beacon of love and support in my life, past and present. I always can count on you to listen and care for me, even when we're so far apart. I love to see how our relationship has changed over time, and can't wait to see what the future holds.

Finally, I want to thank my wonderful partner and surest friend, Dennis Sun. I was incredibly lucky to arrive in Berkeley and be able to learn and work in such an engaging environment, but this lucky break was utterly eclipsed by the most important lucky break of my life in Berkeley: meeting you. We've been through full-hearted laughs together (I'm thinking of a particular scene from Star Wars), stressful trips to the emergency room, a full-on Apocalyptic day of red smoke, and even a handful of fights over dumb data analysis questions. Thank you for being a complement to my distracted optimism, a thoughtful and loving partner, and the best co-cat-dad a guy could ask for. I love you.

Chapter 1

Introduction

The speech signal is characterized by a high degree of variability across different environments, speaking rates, and individual speakers. Understanding how listeners process this variable input signal and arrive at an interpretation of a speaker's intended meaning is a key goal of speech perception research. Experimental evidence has demonstrated that listeners may utilize informative cues in their environment and the acoustic signal to interpret the linguistic message. Social information, such as knowledge of a speaker's age, gender, ethnicity, or region of origin also plays a role, and is able to shift listeners' categorizations of sounds and lexical items. An open question is the degree to which these social cues are integrated and processed like other traditional cues to linguistic contrasts, and to what extent they differ.

This dissertation aims to address this question. To do so, I first survey relevant work on the integration of acoustic, lexical, and social cues during speech perception, and outline several key open questions for the treatment of social cues. Then, I present the results from a series of novel perception experiments aimed at addressing these open questions. Chapter 2 presents the results from a series of selective adaptation experiments investigating the degree to which acoustic cues to voice gender, visual face gender cues, and acoustic cues to perceived male sexuality can shifts American English listeners' perception of fricatives. Chapter 3 presents the results from a perceptual compensation for coarticulation experiment that investigates whether cues to gender can not only induce changes in the perception of consonants, but whether those changes to perception also induce downstream perceptual compensation effects in the following sound. In addition to the classification data from this task, this experiment included on-line eyetracking measurements which allow us to explore the influence of unfolding acoustic cues in real time.

1.1 Phonetic Cue Integration

1.1.1 Acoustic Cues and Trading Relations

Contrasts between sublexical phonetic units are signaled by temporal and spectral characteristics of the acoustic signal. These acoustic characteristics, or PHONETIC CUES, assist in the interpretation of the intended sub-lexical units and linguistic message. A given phonetic contrast is often realized by various overlapping phonetic cues (Polka & Strange, 1985; Raphael, 2005; McMurray & Jongman, 2011), as in the case of voicing contrasts of English stops in medial position, a contrast which recruits more than a dozen phonetic cues (Lisker, 1986). Additionally, the cues recruited to signal a contrast may vary due to differences in phonetic context, speaking rate, and speaker; a one-to-one mapping of phonetic contrast to phonetic cue(s) is untenable, leading to the "lack of invariance" issue (Liberman et al., 1967). Understanding how listeners map variable, multi-dimensional phonetic cues onto phonetic contrasts is a key research goal in speech perception.

When multiple phonetic cues signal the same phonetic contrast, listeners tend to integrate these disparate acoustic cues into a coherent percept, as illustrated by instances of duplex perception (Whalen & Liberman, 1987; Fowler & Rosenblum, 1990). Not all phonetic cues are equally recruited during perception; each cue's relative contribution is weighted (Francis et al., 2008; Idemaru & Holt, 2011). For example, American English /i/-/ɪ/ can be signaled by differences in formant frequencies (/i/ has a higher first and second formant frequency than /ɪ/) as well as duration (/i/ is longer than /ɪ/). Listeners are sensitive to both cues to this contrast, but generally appear to rely more on the spectral cues than the duration cues (Hillenbrand et al., 2000). However, the relative weighting of the duration cue is larger in formant-synthesized stimuli (with ambiguous spectral information) than in sinusoidal-synthesized stimuli, possibly due to the reduced quality of the spectral cues. This result can be interpreted as an instance of TRADING RELATIONS (Repp, 1981), where the relative weighting of two or more cues can change as a function of the informativity and/or reliability of each individual cue. Similar trading relations exist in the perception of listeners with cochlear implants (which involve some degree of spectral degradation of the signal), where these listeners rely more on durational cues than listeners without cochlear implants (Winn et al., 2012).

Trading relations are not only due to changes in signal fidelity, but can also be conditioned by specific language experience. Qualitatively different cue-weighting strategies have been observed when comparing different languages (Beddor & Krakow, 1999; Kang et al., 2016), different dialects of the same language (Lee et al., 2013), individual idiosyncrasies (Massaro & Cohen, 1977; Idemaru et al., 2012; Shultz et al., 2012; Kong & Edwards, 2016), different L1 and L2 listeners (Yazawa et al., 2019), and exposure to different distributions of cues in training data (Holt & Lotto, 2006). There is also evidence that specific cue-weighting strategies develop over time in young children, before eventually settling on adult-like strategies (Slawinski & Fitzgerald, 1998; Nittrouer & Lowenstein, 2009). Cue-weighting strategies may also generalize, such that listeners develop sensitivities to cue-dependencies between related phonetic units. For example, voice onset time (VOT) values across different places of articulation are correlated (Theodore et al., 2009; Chodroff & Wilson, 2018). Listeners are sensitive to this codependency and will generalize it from one context (e.g., long /p/ VOTs) to other unheard contexts (e.g., expecting long /k/ and /t/ VOTs) (Clayards et al., 2008; Nielsen, 2011). Results such as these have motivated models of phonetic cue integration that propose that experience with statistical distributions during language learning, rather than fully innate and psychoacoustic mechanisms, plays a role in the development of cue-weighting strategies in perception (Toscano & McMurray, 2010; Kleinschmidt & Jaeger, 2015).

1.1.2 Intrinsic and Extrinsic Cues

Phonetic cues can be categorized by their locus of realization. For example, the formant and duration cues to vowel identity discussed above occur within the scope of the vowel they contribute to. We can classify such cues as INTRINSIC since they are associated with and occur within specific sub-lexical units. But informative cues to the identity of sub-lexical units come from external sources as well. Because speech sounds are produced using physical articulators which must spend time transitioning from one state to another (Browman & Goldstein, 1992), acoustic boundaries between segments are not discrete, but rather continuous (Ellis & Hardcastle, 2002). The influence of nearby segments on the production and acoustics of a sound is often termed COARTICULATION[1], and can be either anticipatory (e.g., vowel nasalization before a nasal consonant) or perseverative (e.g., vowel rounding following a labial consonant) (Hardcastle & Hewlett, 1999).

Listeners are sensitive to the long-distance effects of coarticulatory forces and use these as EXTRINSIC CUES[2] to the identity of sub-lexical units. This sensitivity is clearly demonstrated in *gating tasks* (Grosjean, 1980), in which listeners hear progressively longer portions of words and are asked at each gate what word or sounds they hear. Listeners encountering a gate can consistently use anticipatory coarticulatory cues to predict what the following material would be (Kuehn & Moll, 1972). This sensitivity is not restricted to immediately adjacent segments, but can occur across intervening sounds (Tobin et al., 2010; Grosvald & Corina, 2012). The informativity of coarticulatory information is further supported by instances of conflicting extrinsic and intrinsic cues, such as when segments produced in one context are spliced into another context. In such situations, listeners' reaction speed and identification accuracy are decreased (Martin & Bunnell, 1981; Whalen et al., 1993; Dahan et al., 2001). This indicates that extrinsic cues are not simply effects observed in production, but key input to perceptual processes.

[1] This usage is distinct from the description of doubly-articulated consonants as 'coarticulated' (e.g., labio-velars).

[2] Though often used to include non-linguistic cues (e.g., speaker identity, sentential context, emotion), I restrict my definition of 'extrinsic cues' to cues originating in adjacent regions of the acoustic signal.

1.1.3 Overlapping Cue Sources

When interpreting the identity of sub-lexical units, listeners must contend with the fact that a given acoustic property may be simultaneously influenced by multiple factors. In the case of acoustic cues, a segment may have a particular acoustic value because it is an intrinsic cue of the intended segment, because it is influenced by adjacent segments, or some combination of the two. For example, lower spectral center of gravity can be an intrinsic cue to the /s/-/ʃ/ contrast in English, but lower fricative center of gravity could also be due to anticipatory lip-rounding caused by a following round vowel. Mann & Repp (1980) find that the perceptual boundary between /s/-/ʃ/ shifts before a round vowel so that more of the lower spectral energy fricatives are classified as /s/ in this context than before non-rounded vowels. This effect is interpreted as COMPENSATION FOR COARTICULATION; listeners are attributing the lowered spectral energy in the sibilant to a following vowel, and shifting their perceptions of the sibilant accordingly. Compensation for coarticulation effects have been found for a wide variety of cues and contrasts (Samuel, 2011, inter alia). Explanations for compensation for coarticulation are varied, but mechanisms are typically couched within three broad theoretical frameworks: (1) gestural perception (Liberman & Mattingly, 1985; Fowler, 2006), (2) auditory/spectral contrast (e.g., Diehl et al., 2004; Sjerps et al., 2019), and (3) inferential statistical models (Sonderegger & Yu, 2010; Kleinschmidt & Jaeger, 2015). Adjudicating between these frameworks is beyond the scope of the current discussion.

Regardless of the specific mechanism involved, listeners can arrive at an interpretation of coarticulation that is different from the speaker's intended message. That is, they may "fail" to compensate for coarticulation from an adjacent segment and instead interpret the acoustic effect as an intended intrinsic cue of the target segment. Unintended parses of coarticulatory effects have been proposed as a possible mechanism for initiating sound change (Ohala, 1989; Beddor, 2009; Garrett & Johnson, 2013), with individuals demonstrating variability in the degree to which they compensate for coarticulation (Yu & Zellou, 2019, inter alia). A prominent example of this type of sound change is tonogenesis arising from consonant voicing contrasts (Kingston, 2011, inter alia). Because of the aerodynamic demands on the larynx during stop voicing, the fundamental frequency (f0) of a vowel is higher following a voiceless stop than a voiced stop (House & Fairbanks, 1953). Although American English listeners are sensitive to both VOT and f0 of the following vowel, the VOT cue is weighted more heavily by these listeners (Abramson & Lisker, 1985). When the relative importance of the following f0 cue is increased, listeners may interpret vowel tone, rather than consonant voicing, as the main cue to the contrast. This is argued to be occurring in modern day Seoul Korean (Kang & Han, 2013) and Afrikaans (Coetzee et al., 2014), where listeners are increasingly relying on f0 vowel contrasts in the production and perception of what were historically consonantal contrasts.

Like the relative weight of intrinsic cues, the strength and scope of extrinsic cues such as coarticulation can vary according to experience. Although English listeners generally weigh intrinsic VOT cues to initial consonant voicing over extrinsic f0 cues, experience with explicit training is sufficient to shift listeners' strategies to recruit extrinsic f0 cues more (Francis et al., 2008). The greater role of intrinsic over extrinsic cues is not absolute, as extrinsic cues have been shown to be necessary and sufficient indicators of other contrasts (Beddor & Onsuwan, 2003), and the exact weighting of intrinsic and extrinsic cues is dependent on specific language experience (Beddor et al., 2002).

1.1.4 Eyetracking and Neurophysiological Studies of Acoustic Cues

Much of the evidence for the integration of acoustic cues comes from careful manipulation of spectro-temporal characteristics of stimuli and subsequent behavioral measures such as categorization. These experiments represent a rich data source, but introduce a large interval between presentation of the key stimulus information and the dependent measure. Without a finer understanding of the processes occurring during stimulus presentation, it is plausible that differences in listener response to different acoustic cues could be due to processing at a later decision stage, rather than low-level perceptual processes. A strategy to address this question is to employ on-line methodologies, such as eyetracking or neurophysiological measurements, to better understand how perception of acoustic cues unfolds in real time.

Eyetracking has been argued to closely represent the processes of lexical activation and competition as they unfold in real time (Magnuson et al., 2003; Clayards et al., 2008), and evidence seems to support the LINKING HYPOTHESIS (Tanenhaus et al., 2000, inter alia), that listeners tend to initiate unprompted eye-fixations towards lexical items as they are exposed to them. Eyetracking methodologies confirm behavioral results that listeners respond to acoustic cues in a gradient, rather than categorical manner (Kong & Edwards, 2016; Zellou & Dahan, 2019). This gradient sensitivity to acoustic cues unfolds in real-time at the sub-lexical level, as listeners' gaze fixations can be influenced by acoustic information as soon as that information is available (Beddor et al., 2013; Mitterer & Reinisch, 2013; Salverda et al., 2014). Importantly, acoustic cues are not available simultaneously; Reinisch & Sjerps (2013) demonstrate that listeners' gaze fixations are influenced slightly earlier by vowel spectral information than by vowel duration, consistent with the observation that spectral information is available early in the vowel, but duration cues unfold over the course of the entire segment. Processing of acoustic cues is not restricted to within-word contexts, as cross-word coarticulatory information has been shown to guide sub-lexical and lexical activation, and subsequent gaze fixations (Gow & McMurray, 2007; Zellou & Dahan, 2019).

Indirect measures of neural activity offer another window into on-line processing of acoustic information. Spectral and temporal cues undergo complex integration and separation at all stages of auditory processing (Eggermont, 2001). Differences in frequency sensitivity and firing rate due to electrophysiological characteristics of neuronal populations lead to separable streams of neural activation for temporal and frequency components of the signal at the earliest stages of processing in the cochlea, auditory nerve, and mid-brain structures (Pickles, 2015, inter alia). Information from the various streams is integrated at several stages of sub-cortical processing, most notably in the Inferior Colliculus where spectral and temporal information from both ears are combined (Pickles, 2012). At the level of the auditory cortex, spectral and temporal information appear to be differentially processed, with the left-hemisphere auditory cortex responding more strongly to temporal aspects of the signal and the right auditory cortex responding more strongly to spectral characteristics (Zatorre & Belin, 2001; Poeppel, 2003; Hackett, 2015).

Within the auditory cortex, neurophysiological measures have demonstrated that the Superior Temporal Gyrus (STG) plays a critical role in the selective processing of speech acoustics (Yi et al., 2019, inter alia). Direct measurements of surface cortical neural activity in the STG have been correlated with selective activity for sub-phonemic features, such as place-of-articulation, voicing, and specific manner cues (Mesgarani et al., 2014). These patterns of selective response to sub-phonemic features develop rapidly, approximately 100-150ms following the onset of the target sound (Mesgarani et al., 2014). Additionally, the existence of selective regions of onset- and sustained-response neural populations in the STG also points towards the distributed nature of processing in this area, as certain aspects of the spectrotemporal signal are decomposed and analyzed in spatially distinct regions (Hamilton et al., 2018). While further research is required in this area, it is likely the sensitivity to acoustic cues observed in behavioral measures are due in large part to on-line processing.

Neurophysiological studies, like behavioral studies, indicate that linguistic experience can modulate attention to various acoustic cues. Escudero et al. (2009) conducted a categorization task and found that L1-Spanish listeners proficient in Dutch relied more on durational cues to vowel contrasts in Dutch than L1-Dutch listeners. In a follow-up investigation, Lipski et al. (2012) record event-related potentials (ERP) from electroencephalography (EEG) with comparable stimuli and listeners. They find that L1-Spanish L2-Dutch listeners show weaker sensitivity to spectral cues (as measured through Mismatched Negativity (MMN)) than the L1-Dutch listeners in the pre-attentive stages of stimulus processing. Similar investigations have demonstrated that differences between L1- and L2-listeners' neural weighting of acoustic cues can be attenuated with increased language experience or task exposure (Peltola et al., 2003; Ylinen et al., 2009). Taken as a whole, evidence of cue-integration in earlier, more on-line measures of processing indicates that these effects are not solely an artifact of later categorizations.

1.1.5 Multi-Modal Integration of Visual Articulatory Cues

While the acoustic signal is the primary modality of speech, many of the physical gestures which produce the acoustic signal also produce predictable visible movements of the lips and jaw. Such visual articulatory cues could in theory provide some measure of additional information about co-occurring acoustic cues and the underlying gestures which serve as a common source. Listeners appear to recruit these informative cues during speech perception, as presenting congruent audio-visual cues improves speech recognition in noise and for unfamiliar accents compared to audio-only controls (Helfer &

Freyman, 2005; Rosenblum, 2008; Xie et al., 2014; Banks et al., 2015; Bidelman et al., 2020). Overt visual cues such as lip-rounding are also critical in guiding learners to language-specific articulatory strategies (Ménard et al., 2013).

When audio-visual cues and acoustic cues are incongruous, the resulting perceptual effects are intriguing. The experiments of McGurk & MacDonald (1976) paired mismatched audio and visual signals, with the face of "ba" pairing with the acoustics of "ga". Rather than preferentially perceiving one channel's stimulus, or alternating between the two, listeners tended to perceptually merge the two channels' stimuli and report hearing "da". This McGurk Effect demonstrates that visual cues can influence perceptual behavior, and has provided a rich test-bed for investigating cross-modal cue integration (Tiippana, 2014, inter alia).

The integration of acoustic and visual cues appears to only occur when each cue could be causally linked to the same source. For acoustic "ba" and visual "ga", visual cues do not clearly rule out a "da" perception and the cue-integration leads to a McGurk percept. But for acoustic "ga" and visual "ba", the visible lip closure is incompatible with "ga" acoustics and listeners overwhelmingly report the non-integrated "ga" (Saalasti et al., 2012; Olasagasti et al., 2015; Magnotti & Beauchamp, 2017). As Tiippana (2014) points out, even McGurk & MacDonald (1976)'s earliest study prefaces this compatibility argument, noting that visual-only "ga" is often already categorized as "da." Multi-modal integration can persist despite small temporal delays between the audio and visual signals, but significant temporal delays cause the integration effect to disappear and perceptual activity relies only on the acoustic cue, as listeners interpret the signals as coming from two distinct speakers (Magnotti et al., 2013).

Visual cues can induce shifts in categorization similar to compensation for coarticulation of acoustic stimuli. Overt cues to anticipatory lip-rounding of ambiguous /s/-/ʃ/ tokens lead listeners to report hearing more /s/, compared to an audio-only condition (Mitterer, 2006). The increasing compensation for coarticulation with visual information may exist only for ambiguous information, however, as non-ambiguous vowel stimuli do not exhibit compensation differences between audio-only and audio-visual conditions (Kang et al., 2016). Another similarity between acoustic-only and cross-modal acoustic and visual integration is the development of specific cue-weighting strategies during development. Young children rely more heavily on the acoustic portion of cross-modal cues than adults (Sloutsky & Napolitano, 2003; Robinson & Sloutsky, 2004) and are less prone to exhibit McGurk effects, not reaching comparable integration to adults until ages 10-12 (Tremblay et al., 2007; Hirst et al., 2018). These differences due to developmental age appear to be particular to the McGurk effect and are not observed for non-speech multi-modal phenomena (Tremblay et al., 2007).

Differences between multi-modal and uni-modal stimuli provide a window into the inner workings of the integration system. Unlike acoustic-only integration, the cross-modal integration of visual and acoustic cues appears to require significantly more processing resources. In a simultaneous perception task and tactile somatosensory task designed to induce high cognitive load, processing costs are observed for audio-visual perception, but not in the audio- or visual-only conditions (Alsius et al., 2007). Integration of audio and visual cues during a dual task experiment is particularly affected for older adults with greater demands on cognitive resources (Gosselin & Gagn, 2011). Eyetracking studies indicate that adults with Asperger Syndrome (which has been argued to impair general multi-sensory integration) show weaker McGurk Effects than paired controls without Asperger Syndrome, while showing equivalent uni-modal perceptual behavior (Saalasti et al., 2012).

Studies of perceptual adaptation also illustrate a possible difference between McGurk Effect percepts and other auditory "illusions." The selective adaptation paradigm (Eimas & Corbit, 1973; Samuel, 2011) demonstrates that repeated exposure to a given stimulus on a continuum can shift the perceptual boundary such that less of the continuum is perceived as the stimulus. As discussed in 1.1.6, top-down lexical influences can restore absent sub-lexical units and bias ambiguous stimuli towards prototypical sub-lexical categories. These lexically-induced percepts demonstrate comparable perceptual adaptation to their non-induced counterparts (Samuel, 1997, 2001). McGurk induced percepts, however, do not appear to cause perceptual adaptation to the fused percept (e.g, "da"), but showed adaptation to the acoustic component (e.g., "ba" and subsequent decrease in "ba" responses) instead (Roberts & Summerfield, 1981; van Linden et al., 2007; Samuel & Lieblich, 2014). In fact, ambiguous audio (/b/-/d/) paired with unambiguous video (/b/ or /d/) can lead to a perceptual recalibration in the opposite direction to selective adaptation (Bertelson et al., 2003; Vroomen et al., 2004, 2007). That is, listeners will be more likely to report hearing the category of the visual adaptor, rather than less likely.

Similarities and differences between McGurk percepts and non-McGurk percepts are also observable in neurophysiological activity during stimulus presentation. Following exposure to incongruous "ba"-audio/"ga"-video, listeners report more "da" than in the audio-only "ba" condition. This difference exists in early neurophysiological measures as well, as audio-only "ba" reported as "da" induces neural activity in the auditory cortex more similar to veridical "da" stimuli than veridical "ba" stimuli (Lüttke et al., 2016). The similarity of activity between McGurk induced "da" and veridical "da" develop during processing, as activity in the auditory, somatosensory, and visual areas for induced percepts initially pattern like simultaneous "ga" and "ba", before resolving into activity like that of the fusional percept (Skipper et al., 2007). Taking advantage of the high degree of listener variability in the realization of the McGurk percepts (Basu Mallick et al., 2015), Pratt et al. (2015) examine neurophysiological activity over the time-course of perception of stimuli which did and did not result in integrated percepts. Using Electroencephalography (EEG) measures, they demonstrate significant differences between successful and unsuccessful McGurk percepts at the earliest stages of stimulus processing, (between 30-200ms after consonant onset[3]). The network of neural regions involved with audio-visual integration is still an area of active research, but initial studies are consistent in demon-

[3] To account for latency of auditory processing and EEG measurements, the authors introduce a flat 300ms normalization; the critical range corresponds to 330-500ms post consonant onset.

strating parallel interactions between areas associated with auditory processing, speech production, and visual processing (Calvert, 2001; Skipper et al., 2007; Bernstein et al., 2008; Hertrich et al., 2009; Pratt et al., 2015).

1.1.6 Lexical Effects in Cue-Integration

1.1.6.1 Ganong Effect

Integration of phonetic cues during perception is mediated not only by elements of the acoustic signal, but also by more abstract information, such as lexical properties of the frame containing the signal. When presented with an acoustic phoneme continuum (e.g., /t/ to /d/), perceptual boundaries can be shifted towards one end of the continuum if the resulting category would lead to a word while the alternate end would result in a non-word. That is, when presented with an ambiguous signal (e.g., "task" to "dask" continuum), listeners are biased to interpret the signal in a way consistent with a lexical outcome (e.g, more "t"/"task" responses); this phenomenon is called the Ganong Effect (Ganong, 1980; Samuel, 2011). This lexical bias on perceptual behavior does not just exist in tasks where an explicit identification is called for, but can be detected indirectly through selective adaptation paradigms (Samuel, 2001; Samuel & Frost, 2015). Additionally, individual differences in the strength of lexical biases involved in Ganong tasks can predict performance on multi-speaker transcription tasks, with those showing high degrees of lexical influence exhibiting lower levels of transcription accuracy in multi-speaker conditions (Lam et al., 2017). This suggests that the degree of lexical bias is in part speaker-specific, rather than simply task-specific.

The strength of the Ganong Effect has been shown to be more robust (a) when uncertainty in the phonetic cues is increased, such as with degraded signal quality (Gianakas & Winn, 2016) or Specific Language Impairment (SLI; Schwartz et al. (2013)), (b) with increased linguistic experience due to increased L2 proficiency (Samuel & Frost, 2015) or age (controlling for hearing loss, Mattys & Scharenborg (2014)), and (c) under conditions in which lexical biases are stronger, such as identification of final segments in longer words (Pitt & Samuel, 2006), lengthened delays between stimulus and categorization (Rysling et al., 2015), or in conditions of greater cognitive load (Mattys & Wiget, 2011).

These behavioral results are not sufficient to definitively conclude that lexical information is recruited during low-level perceptual processes, since they are also consistent with lexical information being recruited only at later decision stages (Norris et al., 2000). To address this question, eyetracking and neurophysiological studies have been carried out during Ganong tasks. Kingston et al. (2016) collect eyetracking fixation measurements during Ganong tasks and find evidence that both acoustic cues to the target segment and cues to the lexical identity of the frame influence fixations to targets as soon as they become available. When presented stimuli on a dunk-*denk or *dush-*desh continuum, for example, fixations on the "U" visual target are significantly higher in the d-nk frame (where it forms a word) than in the d-sh frame where both options are non-words. Critically, this effect emerges rapidly, even before the acoustic onset of the coda nasal. The presence of nasalization in the target vowel is sufficient to rapidly cue lexical information and result in increased fixations to the "U" visual target. The authors interpret these results as evidence against the gradual build-up of lexical information as proposed in the interactive model TRACE (McClelland & Elman, 1986) and instead find support for feed-forward models such as MERGE (Norris et al., 2000).

The case for rapid lexical effects during early perception is bolstered by neurophysiological studies during typical Ganong tasks, which find significant lexical effects on patterns of fMRI (Myers & Blumstein, 2008) and EEG (Noe & Fischer-Baum, 2020) activity during the earliest stages of phonetic encoding in the superior temporal gyri (STG). Since these differences were observed in the STG, which is associated with auditory processing, rather than solely in areas associated with executive function (e.g., left inferior frontal gyrus [IFG] and anterior cingulate cortex [ACC]), the authors interpret their results as evidence for lexical influence on early, low-level perceptual processes occurring in the STG. Gow et al. (2008) further explore this phenomenon, utilizing Granger causal analysis techniques to examine causality in patterns of phonetic and lexical activation. They find evidence for a causal relationship of lexical information from the left superior medial gyrus (SMG) affecting activation levels in the left STG during early stages of processing (280-480ms post-stimulus-onset). Taken as a whole, these investigations argue for the top-down influence of lexical information in the earlier stages of phonetic processing (280-480ms post-stimulus-onset), but not during the earliest stages of phonetic processing (80-280ms post-stimulus-onset).

On the basis of early causal activation from the SMG to the STG, Gow et al. (2008) ultimately argue in favor of interactive models of lexical and phonetic processing. Kingston et al. (2016) critique this interpretation on two grounds. First, they find issue with the focus by Gow et al. (2008) on a subset of causal relationships which change across the critical periods. Specifically, they question the absence of a causal relationship between the LaSTG and the SMG during the earliest stages of processing (80-280ms post-stimulus-onset). This absence is explained by Gow et al. (2008) as indirect activation by the LaSTG of the AG and then in turn the SMG, suggesting that lexical representations in the SMG are activated not directly by low-level acoustic representations, but rather by abstract prelexical units in the AG. While the critique by Kingston et al. (2016) of the focused selection of particular relationships is well motivated, their argument against the results of Gow et al. (2008) on the basis of "large number of causal relationships and their appearances and disappearances" is less motivated. It is unclear why we might expect the causal patterns of activation during perception to be limited or to remain fixed across perception. The transfer of information between functionally linked neural populations has been proposed to be quite rapid, possibly carried out through oscillations of neural activity in the gamma range (30-80Hz) (Bonnefond et al., 2017). While little is known about the specific temporal dynamics of communication between neural populations in the STG and the SMG, the presence of complex, rapidly evolving causal relationships between related regions is not sufficient to discredit the results of Gow et al. (2008)

Second, Kingston et al. (2016) find issue that the GanongMax stimuli (e.g, [S/SH]ampoo or [S/SH]andal) of Gow et al. (2008) do not exhibit the same causal patterns as the non-word (*sampoo, *shandal) or word (shampoo, sandal) stimuli. It is unclear why we might expect the ambiguous GanongMax stimuli to exhibit the same causal relationships as non-words or words, since the intermediate acoustics could induce separate processing pathways and strategies. It is well known that categorization measures are more sensitive to stimuli which occur in ambiguous acoustic regions (e.g., Feldman et al., 2009), and Gow et al. (2008) specifically address this possibility and present evidence of different neural pathways for ambiguous and unambiguous stimuli during explicit phonetic categorization tasks, where unambiguous stimuli rely more on top-down sublexical phonological information (AG), while more ambiguous stimuli rely more on bottom-up phonetic signal information (STG).

As a whole, the eyetracking and neurophysiological studies point towards a rapid integration of both acoustic and lexical information, likely as soon as such information becomes available in the signal. While Kingston et al. (2016) find issue with the interpretation of Gow et al. (2008) and lexical feed-back models generally, I have argued here that both sets of results are not inconsistent with a lexical feed-back model that operates rapidly.

1.1.6.2 Phoneme Restoration

While the Ganong Effect demonstrates that lexical information can guide perception of ambiguous sub-lexical units, lexical knowledge can also bias perception when the elements of the acoustic signal have been completely masked. This PHONEME RESTORATION effect occurs when a relevant sub-lexical unit is replaced with white noise, a cough, or some other non-speech event (Warren, 1970). In these contexts, listeners will report hearing the original phoneme and perform poorly when asked to locate the non-speech masking event (Warren, 1970; Warren & Sherman, 1974; Samuel, 1997). The strength of the restoration is modulated by the acoustic match between the mask and the phone, with fricatives showing greater restoration by a white noise mask than a tone mask, and vice versa for vowels (Samuel, 1981). Neurophysiological measurements of the STG demonstrate that restored sub-lexical units exhibit patterns of activity which are strikingly similar to their non-masked counterparts (Leonard et al., 2016), though differences in time-course and processing pathway show that these restored percepts are not fully identical to natural, uninterrupted stimuli (Shahin et al., 2009; Leonard et al., 2016). Taken as a whole, the phoneme restoration effect points towards an influence of lexical knowledge on early perceptual processes.

1.1.6.3 Structural Properties of the Lexicon

In addition to the categorical presence of biasing lexical items as demonstrated in the Ganong Effect and Phoneme Restoration, continuous properties of the structure of the lexicon can also influence perceptual processes during recognition.

The most notable of these properties is Lexical Frequency, or the relative frequency of exposure to a given lexical item. Despite being subject to significant individual variability and methodological concerns (e.g., base vs. inflected frequency, appropriate corpora measures, co-linear lexical variables, etc.), frequency is perhaps the most well-studied and consistent lexical variable in investigations of speech perception and spoken word recognition (Baayen et al., 2016). Lexical frequency effects permeate the perceptual system, and high-frequency lexical items are consistently responded to more quickly and more accurately in recognition tasks (Howes, 1957; Luce & Pisoni, 1998)

High frequency words also demonstrate greater retention in serial recall tasks when compared to low-frequency controls (Hulme et al., 1997; Roodenrys et al., 2002), which has been interpreted as evidence for these words' greater resting activation levels compared to competitors (McClelland & Elman, 1986; Luce & Pisoni, 1998; Todd et al., 2019). This explanation is supported by neurophysiological evidence that demonstrates greater neural activity during lexical access of low-frequency words (Fiebach et al., 2002; Prabhakaran et al., 2006; Berglund-Barraza et al., 2019).

An early hypothesis for the source of apparent frequency effects was facilitation for words occurring in dense phonological neighborhoods (Eukel, 1980). Phonological neighbors are customarily defined as two words which differ by only one phoneme (Luce et al., 2000). Words with many phonological neighbors are considered items in DENSE PHONOLOGICAL NEIGHBORHOODS, and words with less neighbors as items in SPARSE neighborhoods. More frequent words tend to occur in dense neighborhoods, hence the earlier proposal that apparent frequency effects were due to true neighborhood effects. However, careful studies of the independent roles of lexical frequency and phonological neighborhood density have consistently demonstrated that the two variables influence perception in opposite directions. While high-frequency words tend to demonstrate facilitative effects in recognition tasks, words in high-density neighborhoods consistently exhibit inhibitory effects on the same tasks (Pisoni et al., 1985; Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Dell & Gordon, 2003). Such results have been interpreted as evidence for processes of competition and lateral inhibition between candidate words during recognition, and were critical in the formation of connectionist models of spoken word recognition and lexical access (McClelland & Elman, 1986; Norris et al., 2000; Norris & McQueen, 2008).

Continuous properties of the lexicon such as lexical frequency and neighborhood density can also guide the interpretation of acoustic cues in the signal. Like the lexical bias observed in the Ganong Effect, interpretation of ambiguous acoustic cues in non-word stimuli can be shifted by manipulating neighborhood density, with listeners more readily interpreting the ambiguous cues as coherent with the high-density interpretation (Newman et al., 1997; Boyczuk & Baum, 1999). The influences of frequency and neighborhood density are also observable in neurophysiological data, where their effects are present at rapid time-scales (Cibelli et al., 2015).

1.2 Social Cue Integration

The act of perceiving and producing speech is not carried out in a vacuum, but is carried out in a rich interactional context with complex overlapping goals, sources of information, and constraints. Investigations of the relationship between the social and the linguistic led to the development of the field of sociolinguistics, and careful studies of the correlations between linguistic variants and social characteristics (most commonly more large-scale demographic characteristics such as age, gender, and class (Labov, 2001)) have demonstrated a profound connection between these two domains. Consider, for example, the principle of *ordered heterogeneity*, a systematic relationship between variable linguistic behavior and social characteristics of communities and speakers. This relationship permeates all levels of linguistic structure, from an individual's allophonic conditioning patterns (Labov et al., 2013) to realizations of discourse and syntactic structure (Díaz-Campos & Zahler, 2018).

The existence of a patterned relationship between social characteristics and speech variability does not require that the speech perception system draw upon such a link during processing. In fact, earlier models of speech perception rejected such a connection, instead holding that the variability among speakers was normalized away, in search of "invariant cues" (Liberman et al., 1967). However, a wealth of experimental evidence in the past few decades has demonstrated that listeners do in fact recruit their knowledge of the co-occurrence of social and linguistic cues in their interpretation of the linguistic signal (Drager, 2010; Foulkes & Hay, 2015). Just as listeners rely not just on acoustics, but also on visual articulatory cues in the external environment to interpret the linguistic signal, listeners appear to be able to incorporate other useful and informative cues like the social identity of a speaker to arrive at their likely intended message. Understanding how, when, and under what circumstances social and phonetic cues are integrated during perception is critical to a complete theory of speech perception. In what follows, I survey a variety of experimental evidence from psycholinguistics, sociophonetics, and related areas in order to determine what is known about this process.

1.2.1 Visual Gender Cues influencing Phonetic Perception

The influence of visual gender cues on acoustic cue perception is robustly attested. Like the incorporation of visual cues to consonant place of articulation (McGurk & MacDonald, 1976), listeners appear to use visual cues to speaker gender to guide the interpretation of gender-conditioned phonetic variation. Differences in vocal tract length caused by sexual dimorphism lead to predictable variations in the spectral realizations of certain sounds as conditioned by gender: fricatives produced by women tend to have greater energy in higher spectral components than those produced by men (Jongman et al., 2000; Fox & Nissen, 2005; Munson et al., 2006b), and the vowel formants produced by men are typically lower than comparable productions by women (Kent & Vorperian, 2018). Average pitch differences between men and women are, to some extent at least, due to similar physical differences.

Listeners have been shown to be sensitive to this relationship and recruit visual social cues during perception. For example, when presented with ambiguously gendered fricatives on a /s/-/ʃ/ continuum, listeners in Strand & Johnson (1996) categorized the stimuli as "s" more often when it was presented with a male face than when it was presented with a female face. This result was detected even when participants were instructed to imagine the stimulus was spoken by a man or spoken by a woman. This pattern of results has been consistently replicated (Munson, 2011; Winn et al., 2013), and the strength of this effect has been shown to be dependent on the degree of overtness of the social information, with explicit gendered faces showing the greatest effect, and implicit gendered sentences showing a weaker effect (Munson et al., 2017).

At first glance, this phenomenon might be interpreted as an extension of the visual integration of articulatory cues; perhaps what listeners are attuned to is not the social cues per se, but inferences about physical size and vocal tract lengths. Such an outcome could sidestep the social altogether, and instead maintain that phonetic knowledge is knowledge about articulations and their acoustics. In fact, novel vocal-tract normalization techniques (Johnson, 2020) provide excellent vowel classification after a single exposure to a speaker's vowel, providing a possible mechanism for this size normalization procedure.

However, evidence of gendered differences in production cross-linguistically and in children suggests that the vocal-tract normalization interpretation cannot be the complete explanation. For example, gender differences in /s/ acoustics in German and English persist when vocal tract morphology is taken into consideration (Fuchs & Toda, 2010). Cross-dialectal and cross-linguistic studies also demonstrate considerable variability in the size of gender differences for vowel and consonant acoustics (Johnson, 2005; Stuart-Smith, 2007; Andreeva et al., 2014). While these cross-variety studies do not explicitly control for vocal tract differences, there is not a consistent regional dimorphism explanation that would account for these results. Gendered differences in acoustic patterns have also been demonstrated for young children whose vocal tracts have not yet experienced changes due to the effects of puberty (Sachs, 1975; Lee et al., 1999). These results demonstrate that gender differences in phonetics are due, at least in part, to the effects of abstract social characteristics, and cannot be fully accounted for by physical differences.

1.2.2 Abstract Social Cues influencing Speech Perception

In addition to gender, listeners have also been shown to display sensitivity to visual cues to social characteristics that do not have a clear link to vocal-tract morphology. Matched visual guise experiments have demonstrated that manipulations of perceived class (Hay et al., 2006b), age (Hay et al., 2006b; Koops et al., 2008; Drager, 2011), and ethnicity (McGowan, 2015; Zheng & Samuel, 2017; Gnevsheva, 2018) have the power to influence intelligibility of speech in noise, shift perceptual boundaries between phonemes, or bias lexical identification. These visual manipulations of ethnicity, region of origin, and class are often carried out through changing the background, clothing, or other contextual elements of the visual speaker, while keeping the speaker and audio the same. While effects of gender on speech perception could be at least in part due to inferences about physical properties of vocal tracts, such an interpretation is not possible for more abstract patterns associated with these other macro-social characteristics. Instead, we must hypothesize that listeners build up knowledge of the arbitrary co-occurrence of these abstract social categories and patterns of phonetic variation.

The influence of social cues on perception detailed above include examples of fairly ecologically-valid manipulations. In our daily lives, inferences about social characteristics of individuals are quickly made on the basis of their clothing, environment, and other visual features (Adams & Kveraga, 2015). The influence of social cues on perceptual behavior is also present in other tasks which are less reflective of how social cues are typically communicated in everyday communication. For example, explicit social labeling (e.g., "You are about to hear a speaker from X.") has been demonstrated to shift categorization responses (Niedzielski, 1999; Hay et al., 2006a). Such an effect is present not just for macro-social categories such as gender or ethnicity, but also for more nuanced and locally defined categories. Explicit labeling of locally defined personae and stereotypes ("This speaker has been called a Valley Girl.") can shift listeners' perceptual boundaries, congruent with expectations of that persona's phonetic patterns (D'Onofrio, 2015, 2018).

The efficacy of explicit labeling is mixed, however, as McGowan & Babel (2019) find that listeners exposed to the same voice with two different guises will shift their qualitative evaluations of the speaker, but will not shift their perceptual response to the acoustic signals. This non-effect on perceptual boundaries persists despite listeners reporting they heard two separate speakers. These results are interpreted as evidence for multiple streams of processing for sociophonetic cues, and demonstrate that perceptual processes may not always be affected by overt explicit labeling (see similar arguments in Drager & Kirtley (2016)). Hearkening back to debates over perceptual versus post-perceptual lexical influence, these results suggest some social influences may exist only in later decision stages.

Zheng & Samuel (2017) explore similar concerns over whether effects reported in sociophonetic experiments could be attributed to post-perceptual influences, rather than integration with perceptual processes. Using the selective adaptation paradigm (argued to represent decision-free perceptual processes), they demonstrate that ethnically-marked face guises shift listeners' ratings of accentedness but do not induce selective adaptation of accentedness ratings. These results are parallel to Samuel & Lieblich (2014) finding that certain McGurk induced percepts do not lead to selective adaptation, but lexically restored percepts do.

These results are complicated by traditional measures of on-line processing such as eyetracking, which appear to indicate that social information can be active at the earliest stages of perceptual processing. Visual guises of age (Koops et al., 2008) and gender (Bouavichith et al., 2019) have been shown to influence eyetracking fixations to lexical items, with listeners showing more fixations to lexical items consistent with the social cue.

Similarly, effects of social cues on on-line lexical access have been observed when those cues are introduced via explicit labeling (e.g., "This speaker is a Californian", D'Onofrio (2015, 2018)), or via sentential frames marked for male sexuality (e.g., "My boyfriend told me to look at the X", Bouavichith (2019)).

A possible explanation for the divergent results between selective adaptation and the on-line eyetracking studies could be a role for sufficient experience with a given sociophonetic pattern. Some support for this view is found by Bouavichith (2019), who demonstrates that implicit priming of male sexuality can shift eyetracking behavior towards lengthened sibilants, but only for listeners with a high-degree of experience with gay male speakers. Additionally, the discrepancy could arise due to separate processing strategies for (sub-)lexical categorization and accentedness categorization.

While the eyetracking studies discussed above all demonstrate social cues inducing overall shifts in looks to targets, the critical social cue information is provided well in advance of the acoustic stimuli. The introduction of significant latencies before the acoustic target introduces the possibility that these effects are caused by decision-level adjustments of expectations during the latency period, rather than direct top-down influence on perceptual processes. Evidence for interactivity of top-down social information and bottom-up acoustic cues would be strengthened if eyetracking fixations could be shown to be driven on-line by social cue information presented without latencies.

1.2.3 Bi-Directionality of Social Cue Information

The existence of a systematic relationship between social and linguistic cues leads listeners to recruit social cues in the interpretation of the linguistic signal. Given this relationship, it is logically possible we would observe a bi-directional effect, with listeners recruiting linguistic cues when making judgments of social characteristics. Experimental work has demonstrated such an effect, with acoustic cues being able to shift listeners' judgments of the social characteristics of a speaker. Robust effects of sociophonetic variables have been observed on the judgments of gender and sexuality (Munson & Babel, 2007; Campbell-Kibler, 2011; Levon, 2011; Mack & Munson, 2012; Walker et al., 2014), ethnicity (Purnell et al., 1999; Thomas & Reaser, 2004), and region of origin (Clopper & Pisoni, 2004, 2006; McCullough et al., 2019).

Fewer studies have investigated how the linguistic signal can influence social perceptions on-line. Bouavichith et al. (2019), for example, utilize a visual-world eyetracking paradigm to determine how visual face gender cues (Female-Male continuum) and sibilant production cues (Sack-Shack continuum) interact on-line. In addition to replicating the effect of gendered face cues on sibilant perception reported in Strand & Johnson (1996), the authors also demonstrate that sibilant identity (as primed via lexical frame) was able to shift perceptual judgments of speaker gender on a speeded gender classification task. The existence of low-level effects of context on the perception of social characteristics is perhaps unsurprising; neural studies of face perception have demonstrated consistent effects of visual contexts on face perception in the earliest stages of stimulus presentation (Wieser & Brosch, 2012; Adams & Kveraga, 2015). However, the exact relationship between perceptual processes involved in visual speaker perception and phonetic perception is an open area of research. Further work is necessary to evaluate the possible interaction between early- and late-stage cue-integration and processing during perception of social and linguistic information.

1.2.4 Social Information and Lexical Representations

As discussed previously, listeners' interpretation of phonetic cues can be influenced by lexical factors. We can also observe similar interactions between social cues and lexical factors during perception. For example, lexical activation of age-graded words (e.g., "old" and "young" words) is found to be faster and more accurate if the speaker's perceived age is congruent with word age (Walker & Hay, 2011; Kim & Drager, 2017). This effect is not simply the build-up of expectations about a speaker over multiple exposures in a blocked experimental design, but is present in rapid low-level perception observed in mixed-talker designs (Kim, 2018; Kim & Drager, 2018). Lexical and social interactions are also observed in priming studies such as Szakay et al. (2016), who employed a cross-language lexical priming task designed to compare priming effects between Māori (MR), Māori-Accented-English (ME), and the more standard Pākehā-Accented-English (PE). The authors demonstrate that while both L1 varieties (ME and PE) prime the L2 (MR), the L2 (MR) only primes items in the ethnically Māori L2 voice (ME), not in the White L1 voice (PE). These results cannot be explained on the basis of greater shared phonetic similarity between MR and ME because the Māori items did not share significant phonetic overlap with the L1 translation equivalents in either dialect. A plausible interpretation of these results is that social characteristics shared between stimuli/speakers, such as inferred Māori-ness, can lead to more robust priming effects in lexical access.

The early integration of social concepts in lexical activation processes is also observable in Implicit Association Tasks such as those conducted by Hay et al. (2019). In this paradigm, participants are responding to interspersed face judgments (old vs. young face, female vs. male face) and lexical decision judgments (real vs. fake word) with either left or right hand button responses. The authors observe that for real-word trials there is a facilitation effect for congruency between the social category of the face judgment and the lexical item. That is, participants respond faster and more accurately to socially-linked words (e.g., "young" words, "female" words) when their face-sorting hand is congruent with that social-link (e.g., Right hand - Old facilitates "old" words for the Right hand - Real condition). Priming effects such as these indicate that abstract social information can influence the early stages of processing, and provide support for models of lexical access and representation that highlight the role of social information in this domain Sumner et al. (2014)

Chapter 2

Selective Adaptation and Gender in American English Fricatives

2.1 Introduction

2.1.1 Perceptual or Post-Perceptual Processes?

From our earlier survey, it is clear that social information can under certain conditions influence the behavior of listeners in speech perception tasks. The vast majority of these studies employ methodologies in which participants are presented with the social and acoustic information and are then asked to make some explicit classification or categorization. While observing significant effects of social information on participant behavior in these tasks is consistent with a direct influence of social cues on perception, it is also consistent with an alternative explanation. Namely, a post-perceptual account in which participants (either consciously or sub-consciously) recruit social cues not during early perceptual stages, but rather during later decision-stages of processing. Under this alternative account, perceptual processes would proceed identically, independent of changes to social cues, and apparent effects of social information on behavior would be indirect, affecting only decision processes.

The distinction between an earlier perceptual stage and a later decision stage has been investigated extensively in the context of lexical influences on speech perception. For example, in debates over the nature and time-course of the Ganong Effect (Ganong (1980); in which ambiguous acoustics are more likely to be perceived in ways resulting in words rather than non-words), Samuel (2001) holds that in these tasks lexical information directly influences early perceptual stages. To support their claim, Samuel (2001) (and similarly in Samuel & Frost (2015)) draw upon data from the Selective Adaptation (SA) phenomenon. In what follows we review this phenomenon and associated experimental paradigm and propose a novel set of SA experiments designed to test the role of social cues on early perceptual, rather than later decision stage, processes.

2.1.2 What *is* Selective Adaptation?

The speech perception process is mutable, and can be influenced by the nature of stimuli it is exposed to. A clear example of this mutability is found in the selective adaptation effect, first observed in the context of speech by Eimas & Corbit (1973). In this experiment, Eimas & Corbit (1973) presented listeners with many repetitions of exposure tokens that were clear endpoints of a given VOT continuum (e.g., either [ba] or [p^ha]). Then, they measured how listeners' classification of that continuum changed depending on whether their exposure blocks were composed of [ba] or [p^ha]. After repeated and lengthy exposure blocks hearing the adapting endpoint, participants classified less of the original continuum as belonging to the same category as the adapting endpoint they heard. For example, a listener who was exposed to many repeated instances of the [p^ha] endpoint would tend to classify more of the continuum as [ba].

Explanations of the mechanism behind the SA effect vary. For Samuel and colleagues, SA is the result of acoustic, lexical, and visual cues to phonetic contrasts influencing early pre-decision stage perceptual processes (Samuel, 1997, 2001, 2011; Samuel & Lieblich, 2014). This interpretation is consistent with usage-based or exemplar-type models of sociophonetic knowledge which hold that social information is directly stored in the linguistic representations and actively recruited during the perceptual process (e.g., Sumner et al., 2014). Additionally, this interpretation of SA is consistent with related computational Bayesian models which cast the speech perception process as inference by an ideal-observer (Sonderegger & Yu, 2010). Kleinschmidt & Jaeger (2015) explicitly model the empirical SA data of Vroomen et al. (2007) and find that these data can be accurately accounted for under this computational framework. Alternative interpretations of the SA effect have been presented, such as the original argument of Eimas & Corbit (1973) that SA effects are the result of auditory processes, where repeated exposure to a stimulus may cause feature detectors to fatigue, thus lessening the auditory/neural response to these stimuli during the later classification task. Regardless of one's interpretation of the specific mechanism behind the SA effect, it is generally assumed that this effect arises from processes active at the early stages of perception, rather than later decision-stage processes, and therefore offers a indirect measure of the influences active at that stage.

2.2 Motivation and Experiments

A more comprehensive understanding of the contexts in which information does and does not induce SA behavior in speech perception can provide insights into the nature of the speech perception process. In what follows, I recruit the SA paradigm to explore whether socially-induced percepts can serve as adaptors, inducing SA effects. This parallels the explorations of Samuel (1997), where lexical information serves to bias ambiguous adaptors to be perceived differently, inducing diverging adaptation patterns, and extends it to a new domain of meaning: gender and sexuality. Experiments 1-3 investigate whether voice gender information can influence the SA behavior of ambiguous sibilants, with slightly varying exposure block structures. Experiment 4 explores whether the influence of gender information on sibilant SA is induced by multi-modal visual face gender cues in the absence of voice gender cues. Finally, Experiment 5 investigates SA to a social cue removed from potential confounds of perceived vocal tract length: male sexuality and sibilant SA patterns. Taken as a whole, these experiments provide novel experimental evidence that social cues, like lexical cues, can demonstrate evidence of Selective Adaptation behavior under specific circumstances. I then discuss the implications of these data for our understanding of the SA process, the role of social information in models of linguistic knowledge, and the speech perception process more broadly.

2.3 Experiment 1 - Voice Gender

Our first set of three experiments investigate whether the SA behavior of American English fricatives can be influenced by the perceived gender of the speaker. If so, this would present evidence for the view that social cues such as gender are active at the earlier stages of speech perception that SA is argued to occur at, rather than solely at later decision stages.

Experiments 1-3 approach this question with an identical set of experimental manipulations and stimuli, but have slight differences in the structure of the exposure blocks which will be explored in each experiment's methods section. Given the identical conditions and similar design, Experiments 1-3 share predictions as well:

1. Classification of the 5-step sibilant continuum will be affected by the sibilant-type of the exposure condition. Participants in the canonical "S" exposure conditions will classify less of the continuum as "S", compared to participants in the "SH" exposure conditions. Participants in the intermediate step conditions will show an adaptation effect between the other two sibilant conditions.

2. Classification of the 5-step sibilant continuum will be affected by the gender-guise of the exposure condition. Participants in the "likely perceived man" exposure conditions will interpret more of the exposure sibilants as "S" and therefore classify less of the continuum as "S", compared to participants in the "likely perceived woman" exposure conditions. Participants in the intermediate step conditions will show an adaptation effect intermediate between the other two gender conditions.

3. Next, I predict there will be an interaction between gender-guise and exposure fricative, with the exposure fricative having the greatest effect in the intermediate step gender conditions, when compared to the other exposure gender conditions.

4. Finally, I predict there will be an effect of block, with the influence of the exposure-conditions increasing throughout the experiment as participants become familiarized with the exposure voice and the task.

In what follows, I first present the methods, analysis, and results for Experiment 1. The links to all pre-registrations and experimental materials for this chapter can be can be found in Appendix A.

2.3.1 Methods

2.3.1.1 Stimuli Creation

Onset	**Coda**
sale — shale	lass — lash
seep — sheep	lease — leash
soar — shore	mass — mash
suit — shoot	mess — mesh
sack — shack	bass — bash
sew — show	brass — brash
sock — shock	class — clash
Sue — shoe	crass — crash

Table 2.1: Target items for stimuli bases. Top-half in gray (sale → mesh) were chosen as critical items to be rated by listeners in the norming experiment.

16 single syllable minimal pairs of English /s/ and /ʃ/ were chosen as potential continua bases during later resynthesis. These 32 total items, shown in Table 2.1, were chosen to balance for fricative position (onset vs. coda), absence of non-target fricatives elsewhere in the word, and to maximize vowel variability. Because of the phonological restrictions in place, it was not possible to perfectly balance for word frequency among the items.

8 speakers (4 men and 4 women) from the UC Berkeley Linguistics community were recruited to provide base recordings of the above 32 target items. Speakers were naive to the purpose of the experiments and were speakers of various North American English dialects. Target words were placed in the carrier phrase “They wanted a X again” in order to provide a consistent prosodic framework and to reduce the coarticulatory effects of adjacent segments. Each recording session was carried out in a quiet space of the individuals’ homes or offices and recorded using a Røde NT-USB microphone with a cardioid response pattern and sampled at a rate of 44100Hz.

Following the recording session, target items were manually extracted from the carrier phrases and their intensity levels were scaled using the SCALE_PEAK() function implemented in the PARSELMOUTH Python library (Jadoul et al., 2018), setting the new peak level to 0.8. Additionally, the intensity of nasal-initial items in the carrier phrase was not appropriate for the words in isolation, so the average intensity of each initial nasal was scaled to be half the average intensity of the following vocalic nucleus. Finally, 250ms of silence was appended to the start and end of each item.

The next stage of stimuli creation involves synthesizing continua between a female speaker and male speaker for each item. Initial explorations indicated that two speakers, W214 and M116, produced high quality continua, due to similarities in voice quality and dialect features. For each of the 32 target items, a continuum between W214 and M116's productions of the target items was created using the TANDEM-Straight Morphing Menu (Kawahara & Morise, 2011). Temporal anchors were placed at each phone boundary, as well as between steady-state and transition phases of diphthongs (as determined by visual inspection of the spectrograms). Continua were generated with 9 steps, with step 1 corresponding to the female speaker W214 and step 9 to the male speaker M116.

2.3.1.2 Stimuli Norming

Stimuli norming was conducted to (a) determine the perceived gender of steps along the continua, (b) evaluate the naturalness of stimuli, and (c) validate that items were perceived as the intended lexical item. 16 speaker continua (the top region of Table 2.1) were chosen for norming on the basis of their perceived naturalness to the author, as well as balancing for sibilant position. Steps 1, 3, 5, 7, and 9 of each of these 16 continua were chosen for norming, for a total of 80 norming items. The norming task was carried out on Amazon Mechanical Turk, with participants being linked to an external website hosted on the UC Berkeley linguistics server. The experiment was constructed using the LAB.JS library (Henninger et al., 2020). Participants first completed a questionnaire to determine their eligibility for participation, test their audio, and gather demographic information. In order to participate, participants were required be 18 years of age or older, currently live in the USA, be native speakers of English, and have no history of speech, language, or hearing disorders.

Over the course of the norming experiment, participants heard a random subset of 40 of the 80 norming items and were asked to type the word they heard, rate how natural the utterance was on a scale of 1-7 (1 = extremely natural, 3 = somewhat natural, 5 = somewhat unnatural, 7 = extremely unnatural), rate what gender they believed the speaker to be on a scale of 1-7 (1 = definitely a woman, 3 = probably a woman, 5 = probably a man, 7 = definitely a man), and finally to indicate how old they believed the speaker to be.

Of the five continua steps tested, step 5 was chosen as to be the intermediate gender step as it had an overall participant mean gender rating of 4.6, the closest to the middle of the gender rating scale (4).

Of the 16 lexical bases tested in norming, 12 continua were chosen that maximized naturalness ratings and were consistently heard as the intended target, rather than another word. These are presented in Table 2.3.

With the 36 normed bases in hand (3 speaker gender steps x 12 lexical bases), I next turned to creating the sibilant continua that would be used for both the classification trials and to create the sibilant-conditions of the exposure trials. The sibilant tokens were

Step 1	Step 3	Step 5	Step 7	Step 9
(1.76; 0.82)	(2.57; 1.14)	(4.67; 1.12)	(6.17; 0.81)	(6.43; 0.70)

Table 2.2: W214 to M116 continuum perceived gender ratings in norming experiment. (Mean; Standard Deviation)

Onset	**Coda**
seep — sheep	lease — leash
soar — shore	mass — mash
suit — shoot	mess — mesh

Table 2.3: Target items for stimuli bases.

likely "S"	**intermediate step**	**likely "SH"**	
A	B	C	**likely woman**
D	E	F	**intermediate step**
G	H	I	**likely man**

Table 2.4: Breakdown of the 9 between-subject exposure conditions.

extracted from the onset sibilant of a 15-step continua from F214 ″sack″ to M116 ″shack″, constructed in Tandem-STRAIGHT as above. Steps 3, 8, and 13 were identified as likely to be perceived as ″S″, ″an ambiguous fricative between S and SH″, and ″SH″. These three sibilant steps were then spliced onto each of the 36 bases, resulting in 108 total exposure tokens organized into the 9 conditions outlined in Table 2.4. Finally, the sibilant steps 3, 6, 8, 10, and 13[1] were extracted in isolation for the classification trials, and 250ms of silence was appended to the start and end of each token.

2.3.1.3 Experiment Design

Experiment 1 was also carried out on Amazon Mechanical Turk, with participants being linked to an external website hosted on the UC Berkeley linguistics server. The experiment was constructed using the LAB.JS library (Henninger et al., 2020). Participants first completed a questionnaire to determine their eligibility for participation, test their audio, and gather demographic information. In order to participate, participants were required to not have participated in the above norming experiment, be 18 years of age or older,

[1] Henceforth, steps 3, 6, 8, 10, 13 from the original continuum will be referred to as sibilant steps 1-5 for clarity.

currently live in the USA, be native speakers of English, and have no history of speech, language, or hearing disorders. The experiment took roughly 10-15 minutes to complete and participants were credited $4 to their MTurk worker account as compensation. If participants were determined to be completing the task in bad faith (e.g., randomly responding, responding with a single response, etc.), they were not compensated and their data were destroyed. This exclusion rate was quite low, with approximately 2% of potential participants being rejected. 180 individual participants completed Experiment 1, with 20 participants assigned to each of the 9 between-subject conditions outlined in Table 2.4.

After the demographic and screening questionnaire, participants then began the experiment proper, where they completed 6 classification-exposure trial pairs, and then a final seventh classification trial.

In the classification trial, participants were instructed to indicate if they heard ″S as in sip″ or ″SH as in ship″. Once they begin the classification trial, they immediately heard a sibilant and were required to press the ″d″ key to indicate they heard ″S″ or the ″k″ key to indicate they heard ″SH″. Participants were allowed to repeat the sound as many times as they wished before coming to a judgment. Once they responded to the stimulus, a 500ms fixation cross appeared in the center of the screen before the next classification stimulus was presented. The classification trials involved a randomized order of 20 total sibilants (4 repetitions each of the 5 sibilant steps: 1, 2, 3, 4, and 5). Every classification trial in the experiment involved the same 20 stimuli, but presented in a different random order each trial.

In the exposure trials, participants were instructed that they were about to hear a speaker say several words, and to pay close attention to the words they heard. Each of the 6 exposure trials randomly selected 10 exposure words without replacement from the total set of 60 words (5 repetitions each of the 12 exposure items per condition). Once participants began the exposure trial, they heard one of the 10 exposure items for that trial followed by 1500ms of silence, then a 500ms fixation cross appeared in the center of the screen, and then the next exposure item was presented. At the end of each exposure trial, the participant was brought back to the instructions screen of the next classification trial.

2.3.1.4 Model Structure and Choice of Priors

This analysis deviates from the pre-registered analysis in one substantial way. In the course of data analysis, it became clear that our question and data required a random by-participant slope for sibilant step to capture the incredible variation between individuals categorization curves, as seen in Figure 2.1. Attempts to implement this random effect structure using linear mixed effects models led to insurmountable convergence issues. Because of the great deal of variability observed in the data, I did not feel confident drawing conclusions from models that did not take by-speaker differences in categorization behavior into account.

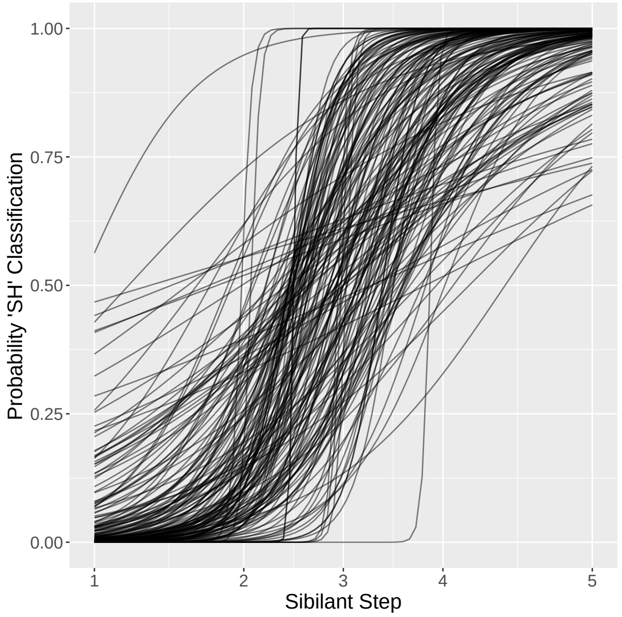


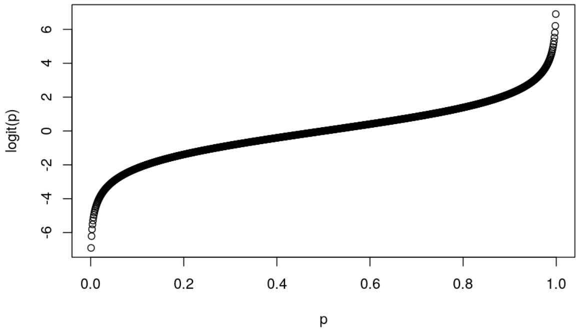
Figure 2.1: Experiment 1 - Classification Curve Variation for Individual Participants by Sibilant Step.

To address this issue, I implement a comparable model in the Bayesian inference framework. In this approach, model parameter estimates are calculated via Bayes' rule, which provides a method for combining prior belief and knowledge about parameter values with the likelihood of observed data given these model parameters. A sketch of this relationship is provided in Equation 2.1.

$$p(model|data) \propto p(data|model) * p(model) \tag{2.1}$$

One benefit of the Bayesian approach over traditional Frequentist methods is practical: this framework does not suffer from as severe convergence issues when fitting more complex models because of the incorporation of both priors and data. A second benefit is theoretical: the ability to constrain the model using (weakly) informative priors allows us to incorporate both the insights generated by previous studies as well as our own expert knowledge. In what follows, I explore the choice of model structure and justify my choice of priors for the first experiment.

The models discussed in this chapter are all logistic regression models: predicting binary outcomes (in this case, choice of “s” (0) or choice of “sh” (1)) by using a logit-link function to transform these binary response probabilities into a continuous logit scale. In this logit-space, the dependent response variable can be modeled as a linear combination of the independent predictors. The logit-link function is presented in Equation 2.2 and visualized in Figure 2.2. A probability of 0.5 of an event occurring corresponds to a logit value of 0. Positive logit values corresponding to increased probability of the event occurring, while negative logit values correspond to decreases in the probability of the event occurring.



$$logit(p) = ln(\frac{p}{1-p}) \tag{2.2}$$

Figure 2.2: Logit-Link Function.

Since the interpretation of the model parameters in logit-space can be non-intuitive, model predictions and/or parameter values may also be transformed back into probability-space using the inverse-logit transformation, presented in Equation 2.3. Special care will be given to indicate when results and parameters are being presented in terms of logit-space or probability-space.

$$inv.logit(x) = \frac{e^x}{(1+e^x)} \tag{2.3}$$

The model specification for experiment 1 is presented in Equation 2.4. We are predicting participants' responses (0 = "S", 1 = "SH") to individual sibilant stimuli in the classification blocks. The model contains a four-way interaction between the main predictors, as well as by-participant random slopes for sibilant step, capturing the individual differences in the effect of sibilant step discussed above. Sibilant step is an ordinal predictor corresponding to the step along the continuum of the classification token (1, 2, 3, 4, or 5). As mentioned previously, these sibilant steps were chosen to cover the range between /s/ and /ʃ/ categories. Experiment block is an ordinal predictor (0-6) corresponding to the block number the specific classification took place in. Recall that classification blocks 1-6 each occur after an exposure block, while classification block 0 is the pre-test block.

$$response \sim sib\_step * block * Exposure\ Gender * Exposure\ Fric. + (sib\_step|participant) \tag{2.4}$$

Recall that the two exposure predictors (Exposure Gender and Exposure Fricative) are between-subject condition manipulations, differing in the nature of the exposure materials. Each of these are modeled as categorical predictors (with dummy coding). Exposure

Parameter	Type	Reference Level	Prior
Intercept	-	-	$\mathcal{N}(-3,\ 1)$
Sibilant Step - *b* Sibilant Step - *simplex*	Ordinal : Monotonic	1	$\mathcal{N}(0,\ 0.5)$ Dirichlet(1)
Experiment Block - *b* Experiment Block - *simplex*	Ordinal : Monotonic	0	$\mathcal{N}(0,\ 0.5)$ Dirichlet(1)
Condition Gender	Categorical	Intermediate Step	$\mathcal{N}(0,\ 0.5)$
Condition Fricative	Categorical	Intermediate Step	$\mathcal{N}(0,\ 0.5)$
Interaction Terms	-	-	$\mathcal{N}(0,\ 1)$
Random Effects Terms	-	-	Half-Cauchy truncated at 0, scale parameter of 0.2

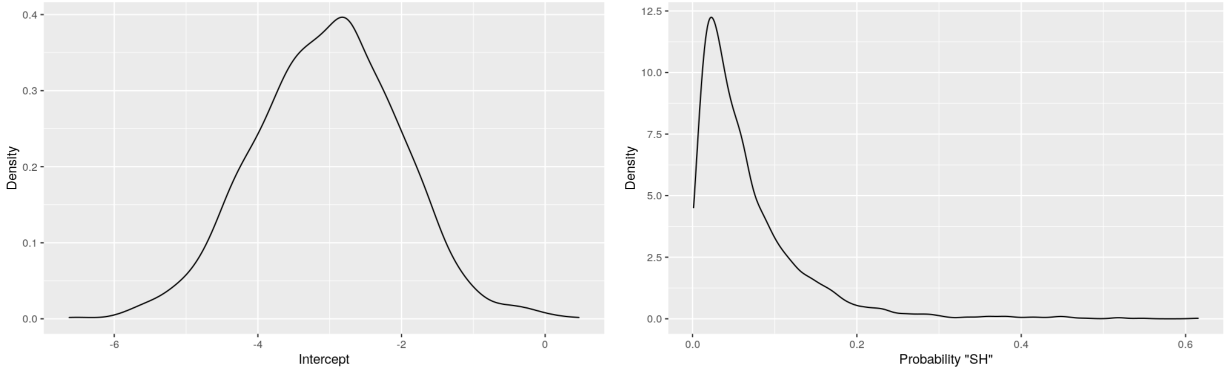
Table 2.5: Priors for Experiment 1 Model.

Gender has three levels: Likely Perceived Woman, Likely Perceived Man, and Intermediate Step voice (reference level). Exposure Fricative has three levels: Likely Perceived "S", Likely Perceived "SH", and Intermediate Step fricative (reference level)

Although it decreases the direct interpretability of model parameter results, the inclusion of a four-way interaction in Equation 2.4 is critical to testing our specific predictions. An interaction between EXPOSURE GENDER and EXPOSURE FRICATIVE, for example, allows each individual exposure condition to be estimated separately; recall that our prediction 3 specially predicts that the effect of the intermediate exposure fricative condition will differ depending on the exposure gender. The inclusion of SIBILANT_STEP in this now three-way interaction allows for the exposure effect to differentially affect the classification continuum, representing our belief that the differences will be most visible in the middle of the sibilant continuum, while the endpoints will be classified at essentially floor and ceiling levels. Finally, the inclusion of BLOCK allows for the various effects to change during the experiment, reflecting our expectation that individuals will shift their classification behavior as they become more practiced in the task and hear more of the exposure condition materials.

Table 2.5 presents the prior distributions chosen for the various parameters for Experiment 1. For all predictor variables, I have chosen weakly informative priors centered around 0. Since these priors are specified in logit-space rather than probability-space, that means that a value of 0 corresponds to no shift (compared to the baseline intercept). Recall that our intercept in this case corresponds to when the ordinal and categorical variables are at their reference level: (sibilant step 1, block 0, intermediate step exposure gender and fricative). Given it is the stimulus step with the highest spectral energy and most [s]-like acoustic properties, we have strong prior belief that sibilant step 1 should be all but categorically classified as "S", regardless of the experiment block or exposure condition. As such, we are able to choose an informative prior of $\mathcal{N}(-3,1)$ for the intercept term. The relationship between this prior distribution and probability-space is shown in Figure 2.3 where 2,000 draws from this prior are presented.

The ordinal predictors (sibilant step and block) are modeled as monotonic predictors, with 2 parameters each: the scale parameter, b, and the simplex parameter, ζ. Following (Bürkner & Charpentier, 2020), I specify a weakly informative prior over the scale parameter centered on zero while still allowing for "large but plausible group differences" (427) as well as a default Dirichlet prior over ζ which corresponds to the assumption that all differences between adjacent groups are equivalent.



(a) Logit-space

(b) Probability-space

Figure 2.3: Visualization of 2,000 Draws from Intercept Prior of $\mathcal{N}(-3, 1)$.

All models in this chapter were fit using the BRM function of version 2.17.0 of the BRMS package (Bürkner, 2021). Arguments were kept to their default values, with the following exceptions: the response distribution family was Bernoulli, the backend was CMDSTANR (v. 2.30.0), and the default 4 chains were calculated across 20 cores using 5 threads per chain. One diagnostic of a well-specified model is overlapping and well mixed chains. Ill-mixed chains, where chains seem to heavily diverge and not overlap on a given value, are indicative of issues with the model or prior specifications. Visual inspections of the chains for the model parameters found them to be well-mixed, indicating that we did not encounter significant issues with divergent transitions, inappropriate priors, or poorly chosen model structures. This is true for all remaining models presented in this chapter.

2.3.1.5 Results

Turning to consider the model results, direct interpretation of model parameter estimates in a logistic regression with a 4-way interaction between ordinal and categorical predictors is extremely unintuitive. As such, I randomly sample from the fitted posterior distribution of parameter estimates and will visualize these draws to better understand the model output. In what follows, all references to "posterior distribution draws" refer to the same random 4,000 draws from the fitted posterior distribution.

Our interpretation of the magnitude of potential differences between groups will be guided by the Highest Density Interval (HDI) measures from these posterior draws. HDI

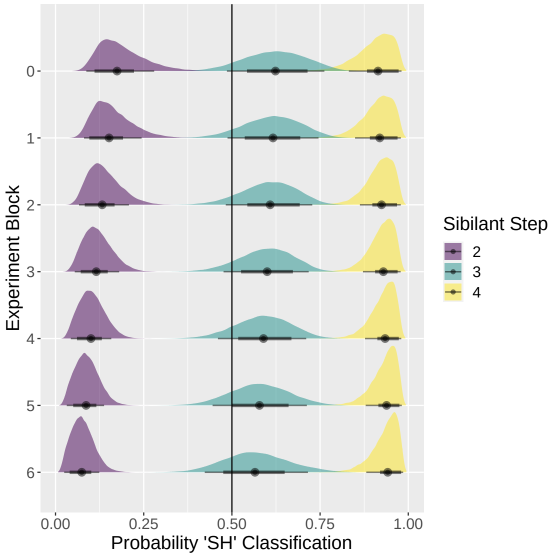
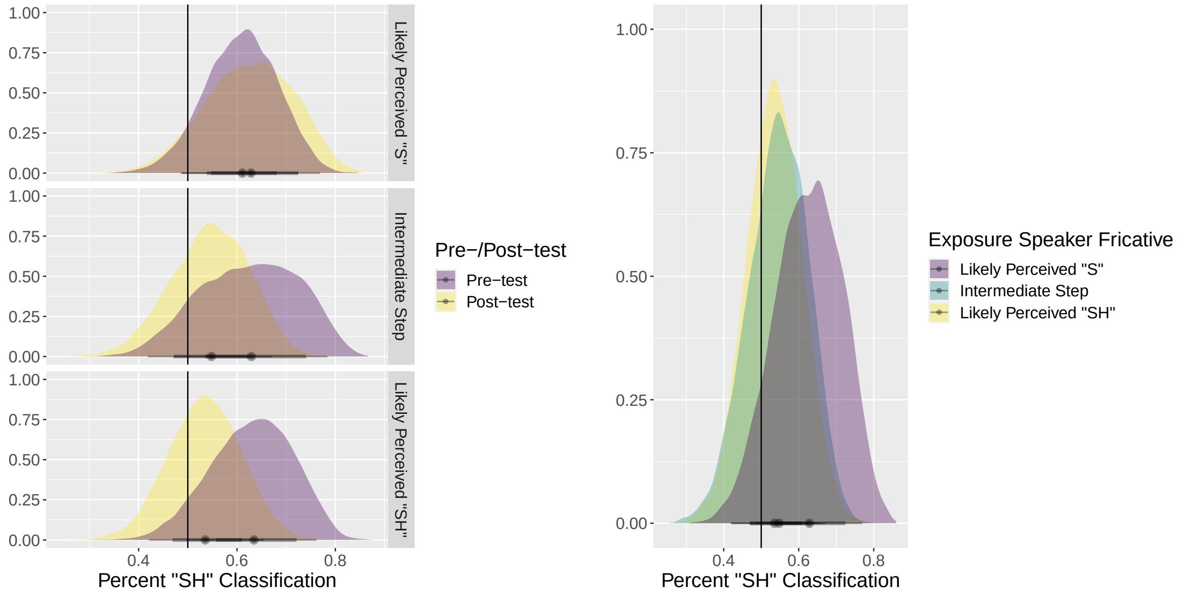


Figure 2.4: Experiment 1 Posterior Distribution Draws: Middle Sibilant Steps (2-4) by Experiment Block.

"indicates which points of a distribution are most credible, and which cover most of the distribution. Thus, the HDI summarizes the distribution by specifying an interval that spans most of the distribution, say 95% of it, such that every point inside the interval has higher credibility than any point outside the interval" (p. 87, Kruschke, 2015). For the visualizations of posterior distribution draws in this chapter, I present two HDIs for each grouping: one at the 66% level (the thicker, shorter black bar) and one at the 89% level (the longer, thinner black bar). Both 89% and 95% HDIs are used to detect differences between groups, Kruschke (p. 184 2015) argues that at least 10,000 samples must be used to calculated a 95% HDI that is accurate and stable, and thus I have chosen 89% as our arbitrary threshold of group difference. In interpreting the results, I may also refer to group posterior differences whose 66% HDIs do not overlap but whose 89% HDIs do overlap as suggestive differences or trends, but these do not represent as robust evidence of difference as non-overlapping 89% HDIs. Additionally, where relevant in the text, I may refer to specific HDI values; I also provide the complete HDI values for all experiments' post-test block of the middle sibilant step classification in Appendix B.

Turning first to the main effects of stimulus step and experiment block, we see that each of these variables have a sensible and predicted influence on probability of "SH" responses. Posterior estimates for stimuli steps 1 and 5 are at floor and ceiling, respectively, and are omitted from Figure 2.4. The intermediate stimuli steps (2-4) are well-separated with no overlap occurring between each step's 89% HDI, even at pre-test block 0. This effect of stimulus step is made even more prominent as the experiment continues. These posterior draws indicate as the experiment progresses, group-level response distributions



(a) Pre-test (block 0) versus Post-test (block 6)

(b) Post-test (block 6) only

Figure 2.5: Experiment 1 Posterior Distribution Draws: classification differences for step 3 stimulus, separated by exposure condition fricative.

become more categorical, with stimulus step 2 shifting towards 0% "SH" classification and stimulus step 4 shifting towards 100% "SH" classification. This pattern of stimulus step and experiment block indicate that, as norming suggested, our sibilant continuum does cover the entire /s/—/ʃ/ range, and that participants are carrying out the classification task as instructed.

With those predicted effects in hand, we now turn towards the predicted effects of exposure condition fricative and speaker gender. Recall that previous selective adaptation experiments have demonstrated that repeated exposure to a phoneme can lead to less of an acoustic continuum being classified as that phoneme. This led to our prediction that participants in the 'likely perceived "SH"' conditions would exhibit less "SH" responses than other exposure fricative conditions. This predicted effect can be weakly observed in Figure 2.5 which presents the posterior distribution draws for the most ambiguous stimulus step (step 3) broken down by exposure fricative condition. The left panel of this figure shows both the pre-test and post-test distributions, indicating a slight decrease in probability "SH" responses for the intermediate and likely "SH" exposure conditions. The difference between post-test distributions is most visible in the right panel of this figure, which shows that listeners in the likely perceived "S" group show the highest probability of classifying step 3 as "SH", while the intermediate and likely "SH" group are less likely to rate this stimulus item as "SH".

This fricative effect is consistent with our predictions and aligns with previous work

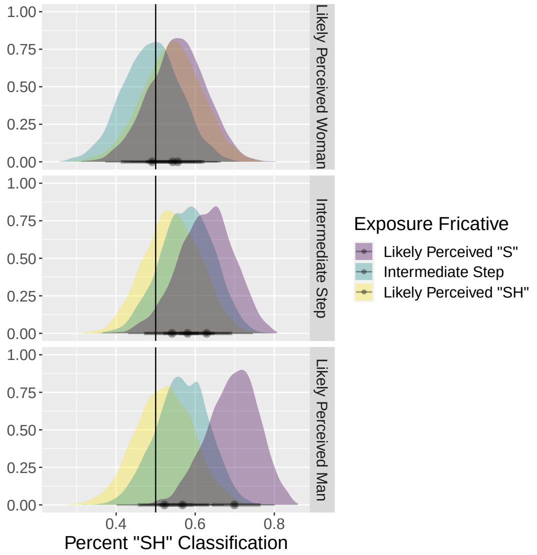


Figure 2.6: Experiment 1 Posterior Distribution Draws: Post-test (block 6) classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

on selective adaptation effects in English. This effect is slight however, with the fricative conditions' HDI (both 66% and 89%) overlapping to an extent, corresponding to a shift in mean percent "SH" classification of approximately 10%.

I turn now to the novel question and predictions of this experiment: the effect of exposure voice gender on selective adaptation behavior. Figure 2.6 presents the posterior distribution draws broken down by exposure fricative and exposure gender, which correspond to the 9 conditions presented in Table 2.4. We can first observe that, averaging across fricative conditions, listeners exposed to the likely perceived woman voice were less likely to categorize the sibilant continuum as "SH" overall. Additionally, considering the interaction between exposure fricative and exposure speaker gender, we see the predicted fricative selective adaptation effect is quite robust for the likely perceived man speaker: listeners presented with this speaker show shifted response patterns to stimulus step 3 depending on which exposure fricative they heard.

This is evident in the "S" and "SH" HDIs for this speaker condition, which have been presented in isolation in Figure 2.7. These HDIs correspond to the HDIs in Figure 2.6 and are interpreted in the same way: the point represents the mean of the posterior draws of that condition, the thick, shorter bar represents the 66% HDI, and the thinner, longer bar represents the 89% HDI. Evidence for the SA fricative effect is found in the HDIs of the likely perceived man speaker, which do not overlap in the case of the 66% HDI ("SH": 0.45-0.59; "S": 0.64-0.77) or barely overlap in the case of the 89% HDI ("SH": 0.40-0.63; "S": 0.59-0.80). For the other two "speakers", this distinction is not as evident as the HDIs

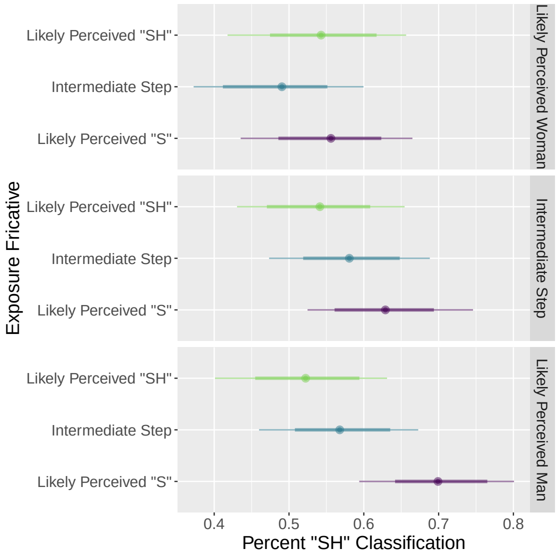


Figure 2.7: Experiment 1 Posterior Distribution Highest Density Intervals: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

for all three fricative conditions overlap to a large extent. However, notice that we do not observe the reverse pattern for either of these two speakers (e.g., *less* “SH” responses after the likely perceived “SH” exposure fricative).

2.3.2 Interim Discussion

The results of this experiment demonstrate a weak fricative SA effect in the predicted direction at the aggregate level (more “SH” responses after “S”). This is driven by a robust SA for the likely perceived man speaker and no evidence of SA fricative differences for the other two speaker gender guises.

One potential source of the attenuated effect compared to other SA studies is the lack of an attention check in our exposure. While participants were instructed to pay attention to the exposure stimuli, they were not required to perform any attention checks during these exposure blocks. It is unclear, especially given the remote and online nature of experimental participation, whether participants truly were attending to the exposure stimuli for all speakers, or rather simply choosing to focus only on the classification blocks.

Another potential cause could come from differences in the structure of exposure blocks. The exposure conditions of other previous SA experiments involve many more stimuli than the current experiment and are quite lengthy. It is possible that a longer exposure block is necessary for the effects of SA to build up and be robustly detected.

2.4 Experiment 2 - Voice Gender: Exposure Block changes and Attention Check

2.4.1 Methods

Experiment 2 is designed to address the methodological differences between Experiment 1 and previous SA studies. The materials and methods are identical to those of Experiment 1, with the following small changes. First, a random attention check was added to approximately 15% of exposure stimuli. In this attention check, participants would be asked to type the word they just heard. These responses were not monitored for accuracy, and served to only ensure participants were actively engaged during the exposure blocks. Second, the exposure blocks were condensed: moving from 6 to 4 exposure blocks. Each exposure block was lengthened as well, moving from 10 exposure items per block to 18. This corresponded to an increase in total exposure items from 60 (5 reps of 12 possible items) to 72 (6 reps of 12 possible items). These changes are designed to increase the salience and amount of exposure stimuli, and detect any effects, if present. Our predictions remain the same as those presented in Section 2.3.

2.4.2 Results

The model-fitting process for Experiment 2 was identical to Experiment 1. Again, given the difficulty of directly interpreting posterior parameter distributions, I instead consider draws from the posterior distribution of all parameter estimates, and present them in probability space, rather than logit space. Additionally, complete HDI values for all experiments' post-test block of the middle sibilant step classification are provided in Appendix B.

Like Experiment 1, we can observe the clear effect of experiment block, with ratings of steps 2 and 4 shifting further towards the 0% and 100% "SH" classification endpoints, respectively. Additionally, the HDIs for steps 2-4 are non-overlapping during all blocks of the experiment, providing evidence that these stimuli steps are reliably distinguished during classification. In Figure 2.8 we observe that posterior distribution of step 3, the most intermediate sibilant step, is extremely wide. This could indicate a uncertain treatment by listeners, or could hide bimodality or structure conditioned by exposure variables.

This is indeed the case, and the wide posterior distribution for step 3 does hide meaningful differences conditioned by exposure variables. Looking first at the aggregate influence of exposure fricative on step 3 stimuli in Figure 2.13, we can appreciate the clear bimodal distributions for the intermediate step fricative and the 'likely "S"' fricative.

This bimodal distribution is caused by the manipulation of exposure speaker gender, as seen in the full exposure condition visualization in Figure 2.10. For the likely perceived woman and likely perceived man speakers, the HDI values (present in Figure 2.10

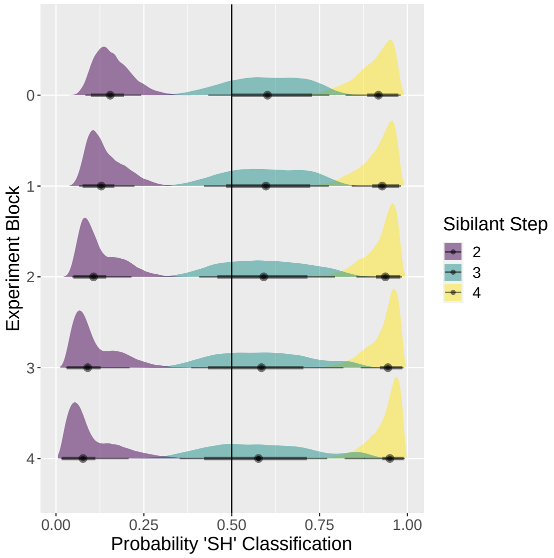
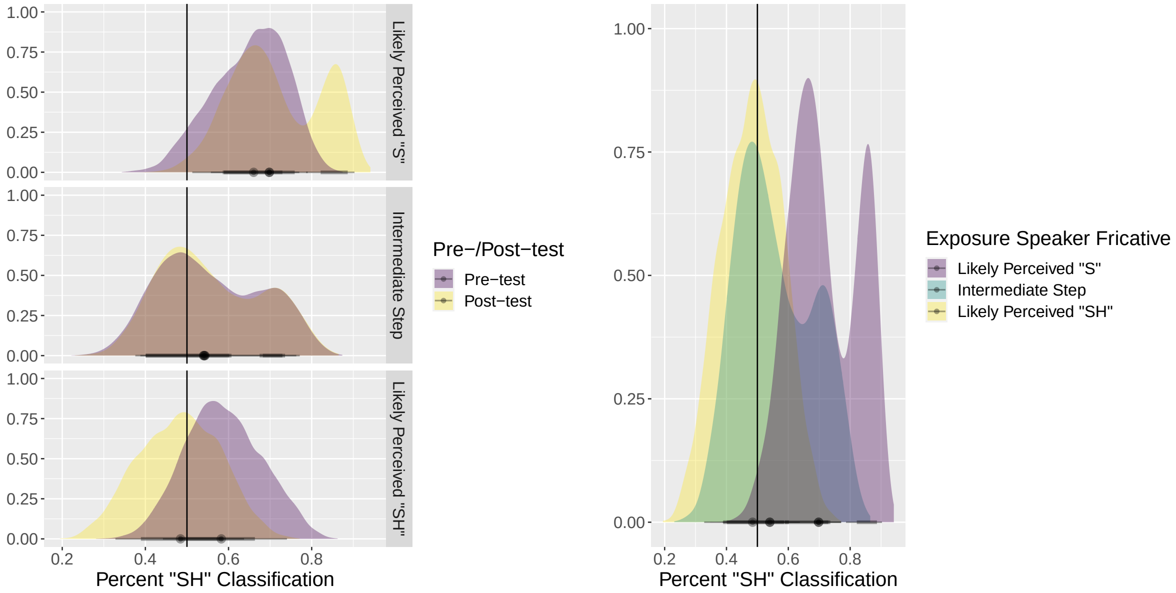


Figure 2.8: Experiment 2 Posterior Distribution Draws: Middle Sibilant Steps (2-4) by Experiment Block.



(a) Pre-test (block 0) versus Post-test (block 4)

(b) Post-test (block 4) only

Figure 2.9: Experiment 2 Posterior Distribution Draws: classification differences for step 3 stimulus, separated by exposure condition fricative.

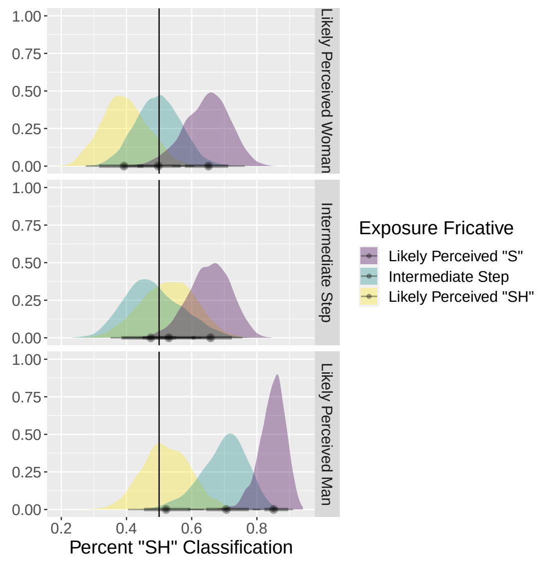


Figure 2.10: Experiment 2 Posterior Distribution Draws: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

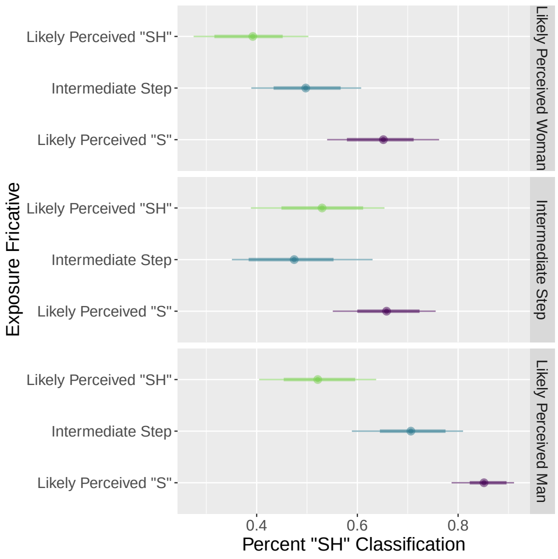


Figure 2.11: Experiment 2 Posterior Distribution Highest Density Intervals: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

and pulled out for visibility in Figure 2.11) for "S" and "SH" exposure conditions are non-overlapping at the 89% level, indicating a high degree of confidence in this difference. The HDIs for "S" and "SH" for the intermediate gender step do not overlap at the 66% level, but do at the 89% level ("SH": 0.39-0.65; "S": 0.55-0.76), which we can interpret as suggestive of a difference. For all three speakers, the difference between fricative exposure conditions occurs in the predicted SA direction: more "SH" responses following clear "S" than when the exposure condition fricative was "SH". We also see tentative evidence for an intermediate SA effect for the intermediate step fricative for the "likely perceived woman" and "likely perceived man" speakers, though the intermediate step fricative appears to overlap with the "SH" exposure fricative for the intermediate step speaker.

2.4.3 Interim Discussion

The methods changes designed to enhance the salience of the exposure blocks (adding an attention check, condensing exposure stimuli into fewer, longer blocks) appears to have achieved the desired outcome. For all three speakers, we observe evidence for the fricative SA effect. This effect develops as the experiment progresses, as seen by the significant influence of experiment block.

We also observe partial evidence for our prediction 2: listeners exposed to the "likely perceived woman" speaker rated less of the continuum overall as "SH" than the "likely perceived man" speaker. The intermediate step speaker, however, did not show an intermediate effect in this case, and instead appears to pattern largely with the "likely perceived woman" speaker.

While the changes to the nature of the exposure block appear to have increased listener attention to the exposure stimuli, this could potentially be due not to an increased SA effect, but rather a confound in the wording of the attention check task. The interpretation of the SA as representative of perceptual processes rather than decision stage processes rests on the assumption that no explicit decision regarding the exposure stimuli is requested. Rather, participants make decisions about other stimuli (in this case the sibilants in isolation in the classification blocks), thus indirectly measuring the effect of the exposure tokens on perception. However, the wording chosen for Experiment 2's attention check ("What was the word you just heard?") can be considered an explicit decision about the exposure fricatives, since that decision is necessary to make an identification of the word it occurs in. I address this potential confound in Experiment 3.

2.5 Experiment 3 - Voice Gender: Updated Attention Check Question

2.5.1 Methods

Experiment 3 was designed to test whether the specific attention check question ("What was the word you just heard?") biased listeners to focus on and classify exposure sibilant. Recall that the argument of the SA paradigm is that the effect of the adapting exposure stimuli is pre-perceptual precisely because its effects are obtained even in the absence of any explicit request to classify the adapting sound. While this attention check is not required an explicit judgment of the adapting sound, it does ask for a decision at the lexical-level. To address this potential confound, I implement a new attention check question which does not call for any explicit categorization of the adapting words. In this new attention check question, participants perform a semantic association task. In the attention checks, participants are asked to type the first other related word that came to mind after hearing the word they just heard. Otherwise, the experiment procedures and methods are identical to Experiment 2. Our predictions remain the same as those presented in Section 2.3.

2.5.2 Results

The model-fitting process was identical to Experiment 2. In what follows, we consider 4,000 draws from the posterior distribution of all parameter estimates, and present them in probability space, rather than logit space. Additionally, complete HDI values for all experiments' post-test block of the middle sibilant step classification are provided in Appendix B.

Looking first at the posterior draws collapsing across exposure conditions, we see in Figure 2.12 a clear and expected effect of sibilant step and experiment block on listeners' classifications. As in previous experiments, the posterior distributions for steps 2 and 4 shift closer to the opposite ends of the classification space as the experiment progresses. Additionally, the HDIs for each sibilant step are non-overlapping at all stages of the experiment.

Turning next to the question of exposure fricative of classification behavior, Figure 2.13 demonstrates the expected SA effect on step 3 categorizations, collapsing across exposure speakers. We observe less "SH" responses to this classification stimulus in the exposure condition "SH" compared to the exposure condition "S", consistent with SA predictions.

This effect persists even when the posterior distributions are separated by exposure speaker gender, as shown in Figure 2.14. Listeners in all three speaker groups demonstrate evidence for differential SA behavior for "S" and "SH" adaptors in the predicted direction: more "SH" classifications of this sibilant step if listeners were exposed to "S" adaptors versus "SH" adaptors. As shown in Figure 2.15, all three speakers' 66% HDIs for

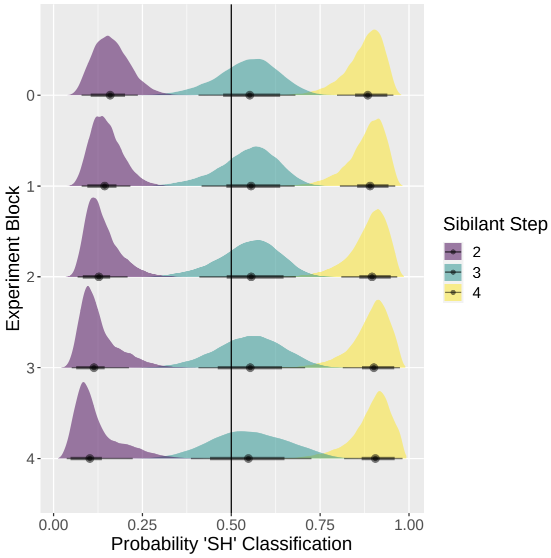
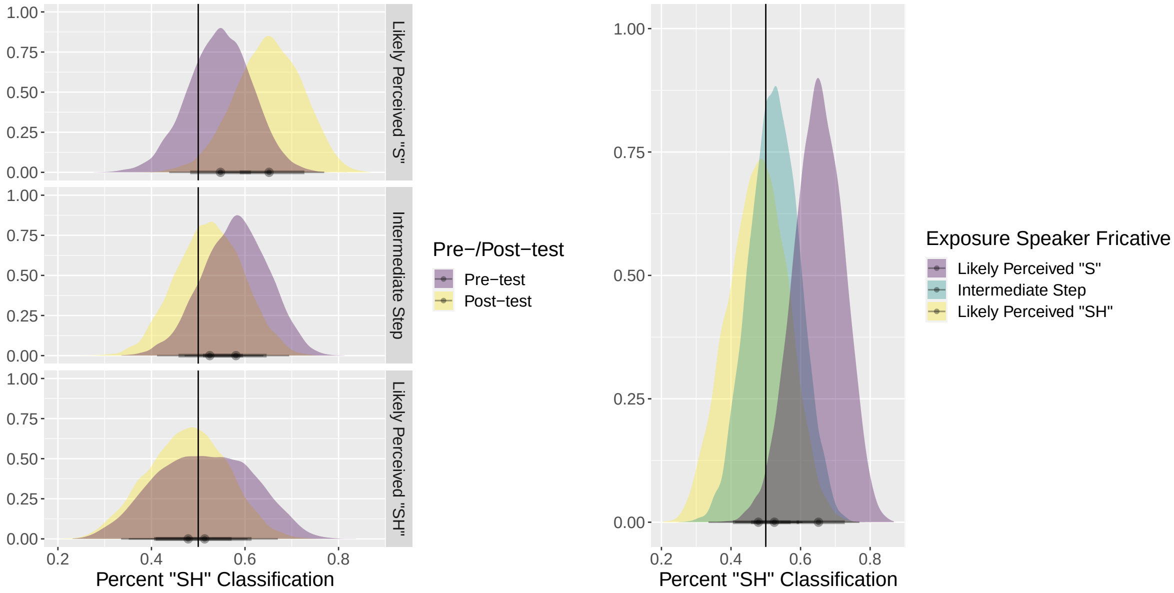


Figure 2.12: Experiment 3 Posterior Distribution Draws: Middle Sibilant Steps (2-4) by Experiment Block.



(a) Pre-test (block 0) versus Post-test (block 4)

(b) Post-test (block 4) only

Figure 2.13: Experiment 3 Posterior Distribution Draws: classification differences for step 3 stimulus, separated by exposure condition fricative.

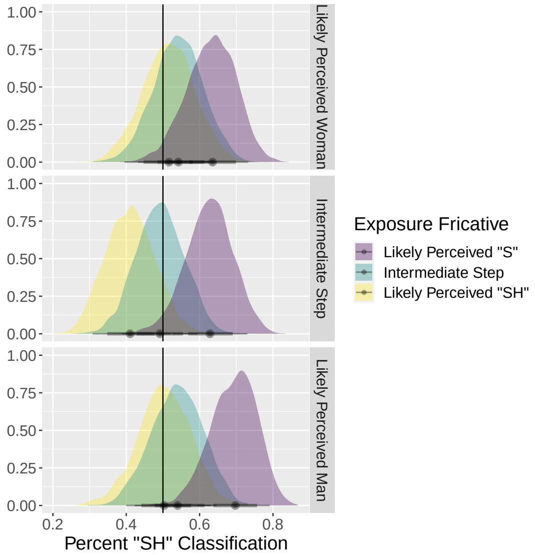


Figure 2.14: Experiment 3 Posterior Distribution Draws: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

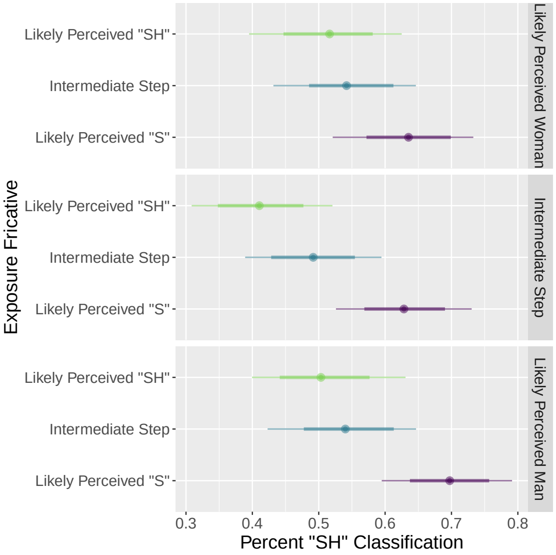


Figure 2.15: Experiment 3 Posterior Distribution Highest Density Intervals: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

"SH" and "S" are non-overlapping or, in the case of the likely perceived woman, nearly non-overlapping ("SH": 0.45-0.58; "S": 0.57-0.70). The 89% HDIs for all three speakers' "SH"-"S" conditions are overlapping, indicating that this potential difference is not as robust as in Experiment 2. Additionally, unlike previous experiments, however, we do not observe differences in the base rate of "SH" classifications between the speaker manipulations.

2.5.3 Interim Discussion

This experiment demonstrates that an altered attention check task (semantic association rather than lexical identification) still induces a Selective Adaptation effect based on exposure fricatives, though this is slightly weaker than the SA effect of Experiment 2. Like Experiment 2, we do not observe evidence for differential SA behavior to the intermediate fricative step conditioned on speaker. That is, we do not find support for our prediction 3, that the speaker gender manipulation would induce differential perception of the intermediate fricative step, and lead to different patterns of SA: patterning with "SH" adaptors for the "likely perceived woman" speaker and with the "SH" adaptors for the "likely perceived man" speaker.

2.6 Experiment 4 - Face Gender

In addition to acoustic cues to a speaker's gender, visual cues through static images or videos of differently gendered faces have been shown to shift listeners' classifications of ambiguous tokens. This effect was first demonstrated by Strand & Johnson (1996), who found that listeners were more likely to classify ambiguous tokens on an [ʃ]-[s] continuum as instances of /s/ if they were paired with a man's face, compared to if they were paired with a woman's face. Listeners appeared to group-level differences in the realization of fricatives, with men's productions of these sibilants exhibiting lowered spectral energy. This visual influence of gender on American English sibilant categorization has been consistently replicated using a variety of materials (Munson, 2011; Winn et al., 2013; Munson et al., 2017). Using a subset of the materials from the previous experiment, this experiment investigates the extent to which visual face gender cues, rather than acoustic gender cues, can induce SA effects.

In a related experiment, Burgering et al. (2020) utilize video recordings of men and women producing Dutch vowels (/e/-/ø/) to determine the extent to which voice gender and face gender cues can induce shifts in vowel categorization.[2] The authors find evidence for a stronger SA effect for the female face and voice than the male face and voice, but the direction of the effect was identical. The gender exposure conditions used by Burgering et al. (2020) contained both unambiguously gendered video cues, but also

[2] Burgering et al. (2020) also investigated how vowel identity influenced gender classification, though we do not explore those results here.

unambiguously gendered acoustic cues around the unambiguous and ambiguous adapting . As such, they are not able to test whether visual face cues alone are enough to induce shifts in classification behavior to sounds when the identity of the adapting sounds is unclear.

If visual cues to social information (present in visual face gender cues) are integrated into the earlier perceptual processes (and not just active at later decision stages), we would predict that face gender cues would be sufficient to guide the perception of ambiguously gendered sibilant tokens. For example, in the absence of clear voice gender information, we would expect an ambiguous sibilant paired with a woman's face to not only be classified more often as /ʃ/ (replicating the classification work mentioned above), but also to shift their SA behavior. If visual face gender can change the interpretation of an ambiguous fricative, we would expect it to also change the direction of its SA effect.

2.6.1 Predictions

This experiment directly tests that prediction, by pairing ambiguously gendered bases with variously gendered faces and exploring how these gender manipulations influence the SA behavior of different fricatives. The predictions for this experiment are as follows:

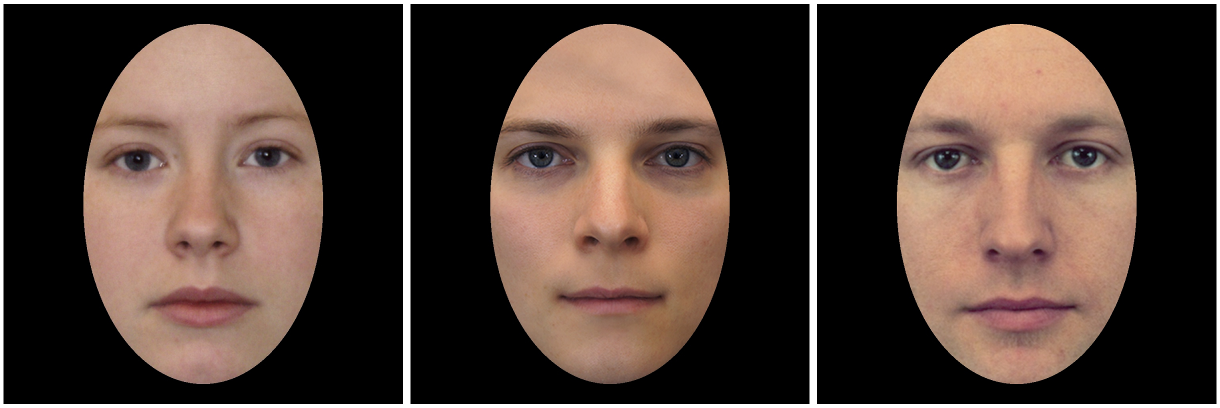
1. Classification of the 5-step sibilant continuum will be affected by the sibilant-type of the exposure condition. Participants in the canonical "S" exposure conditions will classify less of the continuum as "S", compared to participants in the "SH" exposure conditions. Participants in the intermediate step conditions will show an adaptation effect between the other two sibilant conditions.

2. Classification of the 5-step sibilant continuum will be affected by the gendered face guise of the exposure condition. Participants in the "likely perceived woman" face condition will classify less of the continuum as "SH", compared to participants in the "likely perceived man" face condition. Participants in the intermediate step face condition will show an intermediate effect.

3. There will be an effect of block, with the influence of the exposure-conditions increasing throughout the experiment as participants become familiarized with the exposure face and the task.

2.6.2 Methods

The structure of this experiment is identical to Experiment 3, with the following changes to the exposure blocks. First, all exposure stimuli belonged to the intermediate gender step speaker. With the acoustic gender condition removed, I instead implemented a gender manipulation via visually presented faces instead. During the exposure blocks, the face image was presented simultaneous with the exposure stimuli, followed by a fixation cross as in previous experiments. Finally, at the end of the experiment, participants were



(a) Face 1 (b) Face 2 (c) Face 3

Figure 2.16: Synthesized faces used in the visual face gender exposure conditions.

asked to rate on a 1-7 scale (1 = definitely a woman, 3 = probably a woman, 5 = probably a man, 7 = definitely a man) the perceived gender of the person presented in the face image.

The face stimuli used in the present study were developed by Agneta Herlitz[3] and Martin Asperholm[4] and I sincerely thank them for their gracious allowance of their stimuli to be used in this project. All faces were synthesized crosses of real faces, with an oval shaped bounding region to remove hair and clothing cues, and are placed on a black background. For likely perceived woman condition, I chose one of the Herlitz and Asperholm images created by crossing two female faces (here, face 1). For the intermediate face step condition, I chose a face created by crossing a male and female face (here, face 2). Finally, for the likely perceived man condition, I chose a face created by crossing two male faces (here, face 3). These 3 faces are presented in Figure 2.16.

The ratings of each face from participants in Experiment 4 are presented in Figure 2.17. While the ratings for Faces 1 and 3 are relatively uniform and reflect the intended endpoints of this imposed gender continuum, Face 2 is treated more variably.

A summary of the exposure conditions for this experiment is presented in Table 2.6; recall that all exposure bases were that of the intermediate gender step speaker of Experiments 1-3.

2.6.3 Analysis & Results

As in the previous experiments, I eschew direct interpretation of individual parameter's posterior distribution and instead focus on draws from the posterior distribution of all model parameters, yielding posterior density estimates in probability space which

[3] https://ki.se/en/cns/agneta-herlitz-research-group

[4] http://asperholm.xyz

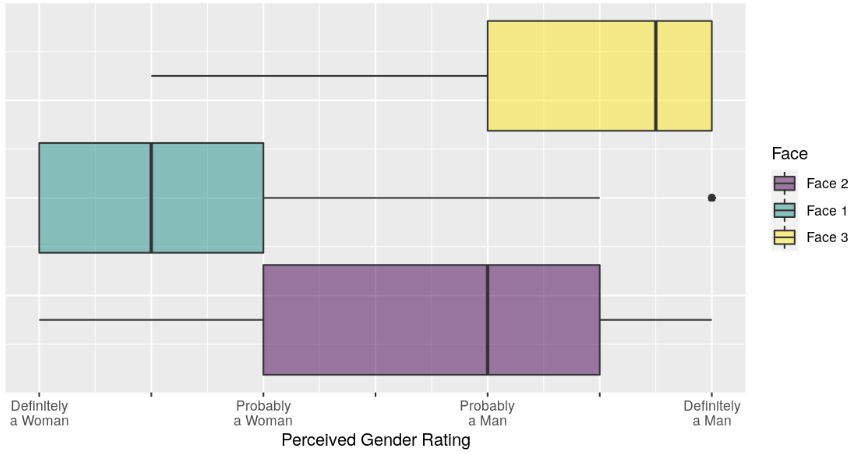


Figure 2.17: Participants' ratings of the perceived gender of each face.

likely "S"	**intermediate step**	**likely "SH"**	
A	B	C	**likely perceived woman face**
D	E	F	**intermediate step face**
G	H	I	**likely perceived man face**

Table 2.6: Breakdown of the 9 between-subject exposure conditions.

take into account each model parameter. Additionally, complete HDI values for all experiments' post-test block of the middle sibilant step classification are provided in Appendix B.

Viewing the posterior distribution draws collapsed across conditions in Figure 2.18, we observe evidence for the predicted effect of block, with steps 2 and 4 moving to towards opposite ends of the classification space. As expected, the sibilant steps 2, 3, and 4 are clearly separated at all stages of the experiment, with their 89% HDIs not overlapping. Additionally, the posterior distribution for step 3 becomes wider as the experiment progresses, perhaps indicative of the emergence of separate distributions conditioned by the specific exposure stimuli.

Turning to the fricative exposure conditions presented in Figure 2.19, we can observe clear evidence for a global fricative SA effect in the predicted direction.

This global fricative SA pattern persists when we analyze each exposure gender condition separately in Figure 2.20. Focusing on the HDIs in Figure 2.21, we see that all three "speakers" demonstrate non-overlapping "SH" and "S" HDIs at the 66% level. The intermediate step face and the likely perceived man face also have non-overlapping 89% HDIs for "S" and "SH" exposure conditions, while the likely perceived woman face overlaps to

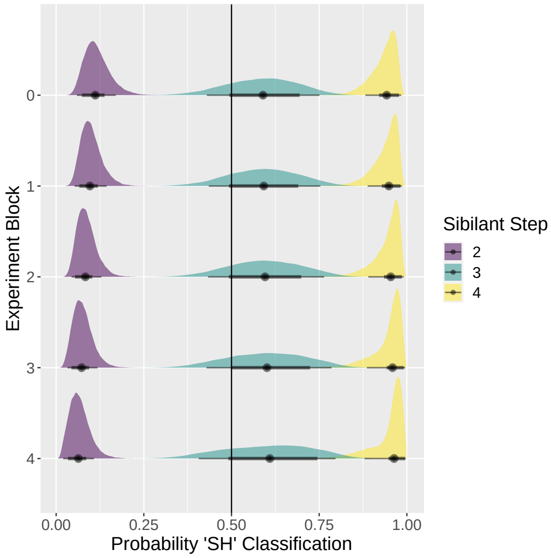
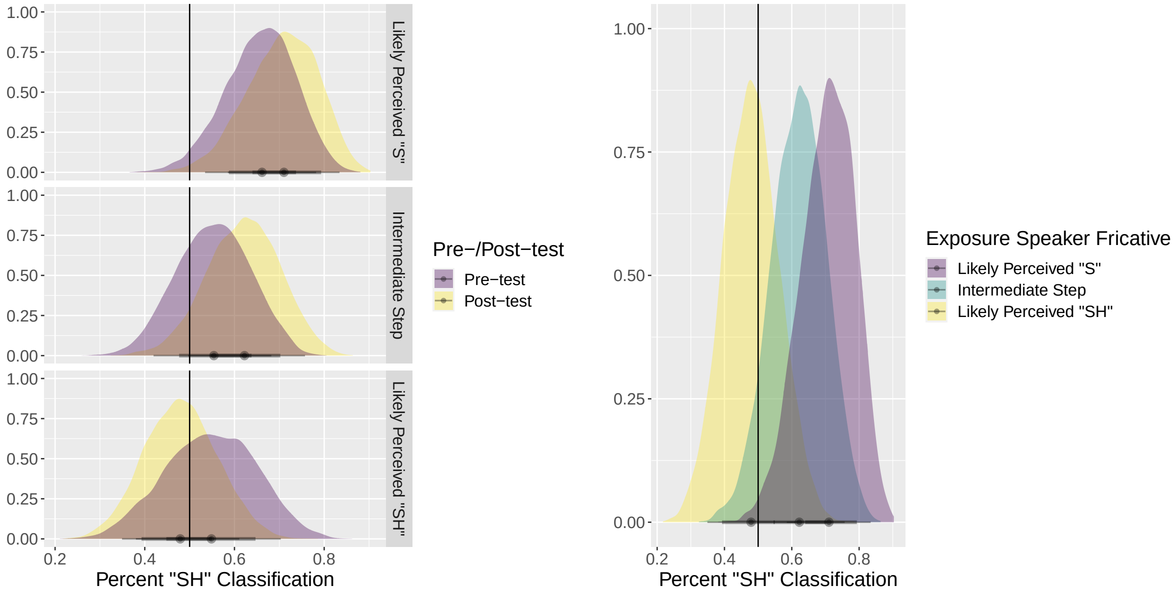


Figure 2.18: Experiment 4 Posterior Distribution Draws: Middle Sibilant Steps (2-4) by Experiment Block.



(a) Pre-test (block 0) versus Post-test (block 4)

(b) Post-test (block 4) only

Figure 2.19: Experiment 4 Posterior Distribution Draws: classification differences for step 3 stimulus, separated by exposure condition fricative.

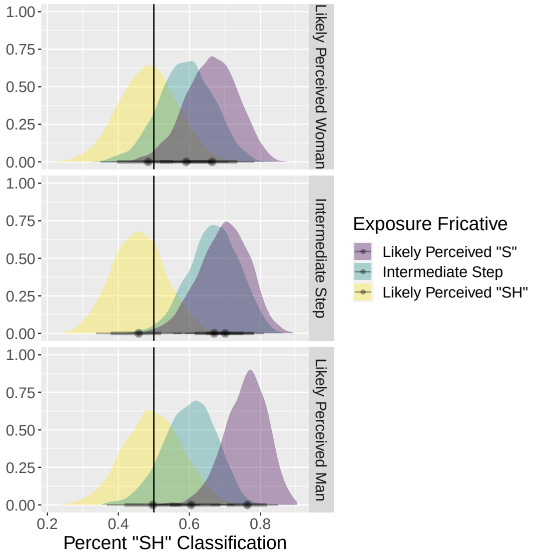


Figure 2.20: Experiment 4 Posterior Distribution Draws: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker face gender.

a small degree (“SH”: 0.35-0.61; “S”: 0.55-0.78).

2.6.4 Interim Discussion

We find clear evidence for the most robust effects of previous experiments: the exposure fricative inducing SA effects in the predicted direction and this effect developing over time as measured via experiment block. However, we observe no evidence for differences between classification in the three face gender conditions. This is despite participants’ clear post-experiment classification of these faces in line with the intended gender manipulation.

Unlike Burgering et al. (2020) who found a fixed effect of gender on the amount of SA observed for Dutch vowels, the size of the SA effect in the current experiment does not appear to be. This is perhaps unsurprising, as unlike Burgering et al. (2020), I present gender information only via the visual face cues, not via acoustic cues to gender as well. It appears that in our case, the presence of static face images was not sufficient to overcome the ambiguous nature of the gender information presented in the voice and our prediction 2 was not supported. It appears that we do not have evidence for a SA effect induced by visual gender cues.

Despite both visual and acoustic gender cues being shown to shift listeners’ categorization behavior, in our data the SA effect only occurs with acoustic gender cues. This highlights a potential dichotomy between the perceptual stages each of these cues are

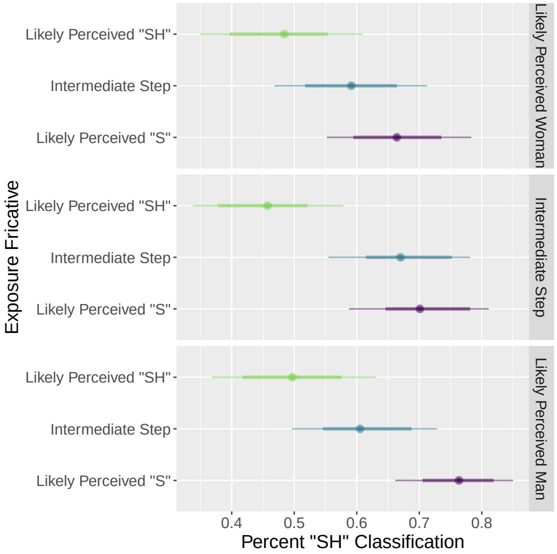


Figure 2.21: Experiment 4 Posterior Distribution Highest Density Intervals: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker gender.

most active in. Under the assumption that SA reflects the integration of information at earlier perceptual stages, we would then conclude on the basis of these experiments that only voice gender cues directly influence these early perceptual processes, with the influence of visual gender cues being found later on and only detectable via the classification behavioral measures.

2.7 Experiment 5 - Perceived Male Sexuality

The influence of gender on the perception and classification of American English fricatives is well studied, and thus represented a natural starting point for the replications and extensions carried out in Experiments 1-4. In the next experiment, I explore the SA behavior for another well studied sociophonetic variable: fricative realization and perceptions of male sexuality.

The relationship between male sexuality and fricative realizations has been well documented, with increased spectral energy, spectral skew, and longer durations of sibilants all being shown to increase the likelihood of a speaker being perceived as gay (Munson & Babel, 2007; Campbell-Kibler, 2011; Levon, 2011; Mack & Munson, 2012; Walker et al., 2014). Higher frequency sibilants are found to be recruited stylistically by both in- and out-group members in the creation and indexing of “gay identity” in American English (see, e.g., Podesva, 2006; Levon, 2011).

Given the relationship between sibilant realization and perceptions of sexuality, it raises the question whether listeners utilizes cues to sexuality during sibilant perception. To our knowledge, only one study has tested for this reverse relationship: Munson et al. (2006a) find no evidence for an influence of perceived sexuality on the categorization of a /s/-/ʃ/ continuum, though their experiment included only 10 listeners and there has not yet been (to our knowledge) an attempt to replicate this experiment. As such, I still consider it fruitful to explore the degree to which this influence (perceived sexuality on fricative perception) may exist.

While modern formulations of exemplar-based models hold that all meaningful social information is included in the multi-modal structure of phonetic representations, can we observe the influence of all such social information on the SA behavior of these sounds? If not, is it possible to articulate a set of criteria that would predict when certain social information would and would not induce SA behavior? Like gender, male sexuality represents a domain in which perception of a social cue is linked to the perception of sibilants. Testing whether we observe SA effects in this new domain can thus help us better understand the scope of social cues which potentially influence early perceptual processes.

2.7.1 Predictions

The predictions for this experiment are as follows:

1. Sibilant categorization will be affected by the sibilant-type of the exposure condition. Participants in the canonical “S” exposure conditions will classify less of the continuum as “S”, compared to participants in the “SH” exposure conditions. Participants in the intermediate step conditions will show an adaptation effect between the other two sibilant conditions.

2. There will be an effect of block, with the influence of the exposure conditions increasing throughout the experiment as participants become familiarized with the exposure voice and the task.

3. Sibilant categorization will be influenced by the sexuality guise of the exposure condition. Participants in the “likely perceived straight” exposure conditions will interpret more of the exposure sibilants as “S” and therefore categorize less of the continuum as “S”, compared to participants in the “likely perceived gay” exposure conditions.

2.7.2 Methods

The experimental design, materials, and analysis of this experiment is identical to Experiment 3, with the following exception: instead of including an exposure condition of speaker gender via acoustic cues in the bases of the adapting words, I introduce an exposure condition of perceived speaker sexuality. To do this, I chose two self-identified male speakers from the set of potential voices that were originally normed, and identified two speakers which were likely to be perceived as straight and likely to be perceived as gay as the endpoints of our resynthesis continuum. The "likely perceived straight" speaker is the same speaker as our "likely perceived man" endpoint in experiments 1-3, though due to the nature of the resynthesis process there may be minor variations in the nature of this speaker's endpoint across continua.

With the two endpoint speakers in hand, I created a resynthesized continuum using the TANDEM-Straight Morphing Menu (Kawahara & Morise, 2011) in a method identical to that described for experiments 1-4. The fricative tokens used for the classification blocks and the exposure fricatives are identical to those used in the previous experiments and are splice onto the new bases.

2.7.3 Analysis & Results

As in the previous experiments, I eschew direct interpretation of individual parameter's posterior distribution and instead focus on draws from the posterior distribution of all model parameters, yielding posterior density estimates in probability space which take into account each model parameter. Additionally, complete HDI values for all experiments' post-test block of the middle sibilant step classification are provided in Appendix B.

Observing the posterior distribution draws in Figure 2.22, we find evidence for a familiar pattern: a clear separation of sibilant steps that becomes more pronounced as the experiment progresses, with steps 2 and 4 moving towards 0% and 100% "SH" classification, respectively. The HDIs for all three steps are non-overlapping throughout the experiment. Additionally, the posterior estimates for the intermediate step 3 cover a wide range of classification space, likely the result of exposure condition based differences.

This is indeed the case, as we observe in Figure 2.23 clear evidence for a fricative SA effect in the expected direction, with "S" exposure conditions falling above the other exposure fricative conditions and inducing more "SH" classifications of this intermediate sibilant step.

Turning to the speaker-based differences in Figure 2.24, classification of this step appears to be uninfluenced by exposure speaker sexuality for the intermediate step and "S" exposure fricatives and the HDIs for these conditions are effectively identical. However, we do find that the SA effect is larger for the likely perceived gay male speaker, with the 66% HDI for "SH" exposure tokens for this speaker (0.26-0.38) not overlapping with the other two speakers (likely perceived straight: 0.39-0.53; intermediate step: 0.50-0.64). At the 89% HDI level, this speaker's "SH" HDI (0.22-0.42) still does not overlap with the intermediate step speaker (0.46-0.48), and only overlaps to a small degree with the likely perceived straight speaker (0.35-0.57).

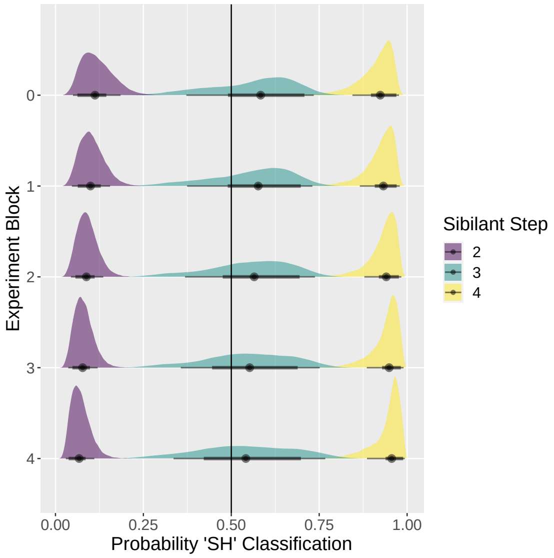
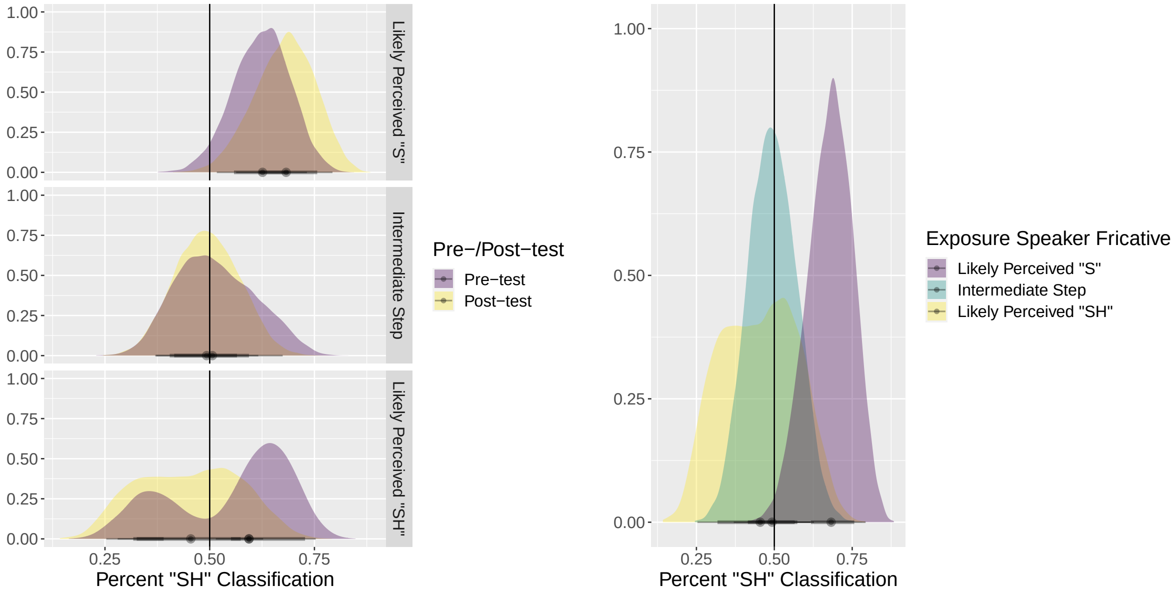


Figure 2.22: Experiment 5 Posterior Distribution Draws: Middle Sibilant Steps (2-4) by Experiment Block.



(a) Pre-test (block 0) versus Post-test (block 4)

(b) Post-test (block 4) only

Figure 2.23: Experiment 5 Posterior Distribution Draws: classification differences for step 3 stimulus, separated by exposure condition fricative.

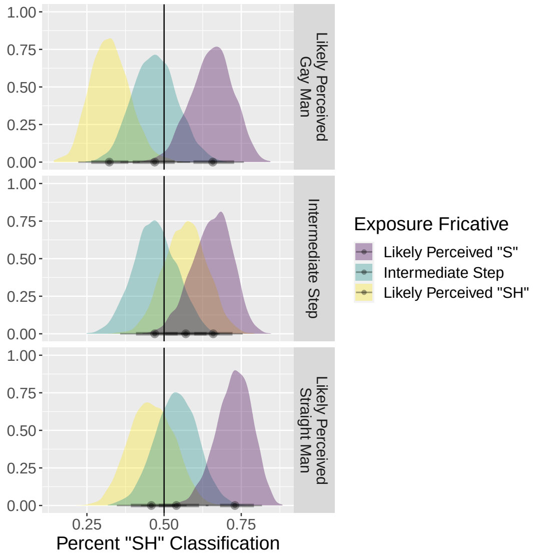


Figure 2.24: Experiment 5 Posterior Distribution Draws: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker sexuality.

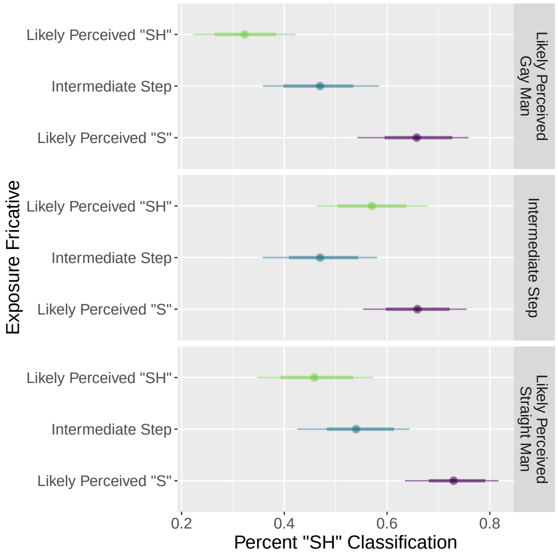


Figure 2.25: Experiment 5 Posterior Distribution Highest Density Intervals: Post-test classification differences for step 3 sibilant, separated by exposure condition sibilant and speaker sexuality.

2.7.4 Interim Discussion

This experiment has replicated a robust selective adaptation effect for fricatives for a new set of listeners and speakers. Our original prediction, that we would observe increased “SH” categorizations for the likely straight speaker compared to the likely gay speaker was not supported overall. However, we do find robust evidence for a difference in the magnitude of the SA effect of the “SH” adaptor between the voices, with this effect on categorization of the intermediate step fricative being extremely robust for the likely perceived gay voice. While this effect was not predicted, it could potentially be explained by listeners’ sexuality based expectations for fricative realizations. If listeners expectations of the likely gay voice are higher spectral energy fricatives overall, the pairing of the “SH” exposure fricative with its dramatically low center of gravity could potentially enhance its perceptual status, causing its SA effect (reducing “SH” categorizations after such an extremely obvious “SH”) to be even further enhanced.

2.8 Chapter Summary and Discussion

This chapter has presented results from 5 separate Selective Adaptation experiments investigating if and how social information can influence the SA behavior of fricatives in American English. Experiments 1-3 demonstrated that robust fricative Selective Adaptation effects were observed for the current set of materials, but only when an explicit attention check was added to the exposure block. Turning to the question of influence of social cues on the SA behavior, we find in Experiments 2-3 evidence for subtle speaker (or gender) based differences in the strength of the fricative SA effect. We do not, however, observe evidence for our core prediction: that the SA behavior of an intermediate fricative step could be shifted to align with the SA behavior of “S” when paired with a likely perceived man’s voice and “SH” when paired with a likely perceived woman’s voice. This pattern continued with Experiments 4 and 5, where we find evidence for a robust fricative SA effects but no evidence of visual face gender cues or perceived male sexuality being able to induce the differential SA behavior in the intermediate fricatives.

Unlike lexical information, which in previous work was found to be sufficient to cause ambiguous sounds to induce SA effects comparable to unambiguous sounds, the social cues explored in this chapter did not shift the SA behavior of ambiguous sounds. Thus, it appears that our main hypothesis of the central role of social cues during the early perceptual processes involved in selective adaptation is not supported.

Chapter 3

Perceptual Compensation for Coarticulation — Eyetracking data

3.1 Introduction

In speech production, individual speech “segments” are rarely if ever produced in an easily segmentable fashion; rather, the articulations of adjacent sounds can overlap and influence one another, leading to patterns of coarticulation. Listeners appear to be sensitive to the patterned coarticulatory relationships between sounds in their language and draw upon these patterns during perceptions. For example, listeners in the experiments of Mann & Repp (1980) classified ambiguous fricatives from an /ʃ/-/s/ continuum more often as /ʃ/ when they preceded an unrounded vowel than when the preceded a rounded vowel. This corresponds to listeners attributing the lowered spectral energy of the fricative to the coarticulatory influence of the following vowel. This effect is known as perceptual compensation for coarticulation (PCCA) and has been robustly attested for a wide variety of segments and languages (Sonderegger & Yu, 2010; Samuel, 2011, inter alia). PCCA effects are not only found in the perception of vowels, but can also be found in adjacent consonants. Mann (1980) introduces a well-known set of PCCA inducing consonant pairs: /l/-/ɹ/ and /d/-/g/. One of the key acoustic features distinguishing /ɹ/ from /l/ is its lowered F3. Another source of potential lowered F3 is the presence of a velar consonant, like /g/, which will induce a so-called “velar pinch” of F2 and F3. Mann (1980) demonstrated that when asked to classify stimuli that were ambiguous between /da/ and /ga/, listeners were less likely to categorize the stimuli as examples of /ga/ when they followed /aɹ/, as they apparently perceptually compensated for the lowered F3 and attributed it to the preceding rhotic, rather than the stop.

Critically, PCCA effects have been demonstrated to be sensitive to information beyond the acoustic cues present in the signal. Elman & McClelland (1988) demonstrate that it is possible to reverse the direction of PCCA caused by an ambiguous fricative on a following stop by using lexical information to change listeners’ interpretation of the ambiguous fricative. For example, when placed in an /s/-biasing lexical context (e.g, "Christma**s** [t/k]apes"), an ambiguous fricative behaved like a clear /s/ and induced a PCCA effect on a following ambiguous /t/-/k/ token, eliciting more "K" responses. When that same ambiguous fricative was placed in an /ʃ/-biasing lexical frame (e.g, "Briti**sh** [t/k]apes"), it elicited a PCCA effect in the opposite direction, resulting in more "T" responses for the following stop.

PCCA effects thus offer an additional method of addressing our core theoretical question on the nature and timing of the influence of social cues on speech perception. If, like lexical information, social cues are able to influence the direction of the PCCA effects, this would be evidence for their central role in early perceptual processes. In what follows I present the predictions for and results from an eyetracking PCCA study and outline the implications of these results for our understanding of the role of social information during speech perception.

3.1.1 Predictions

This experiment investigates PCCA effects for two classes of stimuli: an *asta-ashka* class and an *alda-arga* class. From this point on, I will refer to the first consonant in a pair of adjacent consonants as C1, and the second as C2. As shown in Table 3.1, each class consists of 27 unique combinations of C1, C2, and Speaker. A detailed discussion of the stimuli creation process can be found in Section 3.2.1.

Class	**C1**	**C2**	**Speaker**	**Total**
asta-ashka	S – ? – SH	T – ? – K	W – ? – M	27
alda-arga	L – ? – R	D – ? – G	W – ? – M	27

Table 3.1: PCCA stimuli breakdown. "?" steps represent intermediate steps drawn from the exact middle of the synthesis continuum of the endpoints. For the Speaker column, "W" refers to the "likely perceived Woman" endpoint speaker, and "M" refers to the "likely perceived Man" endpoint speaker.

3.1.1.1 Categorization Data

Let us first consider the predictions for the categorization data from this experiment.

1. **C2 Effect**: more "K" categorizations for "K" C2 steps, more "G" categorizations for "G" C2 steps. This would indicate that our C2 continuum synthesis was effective.

2. **PCCA C1 Effect**: we will observe more "K" responses following "S" than "SH", and more "G" responses following "L" than "R". This would replicate past work on PCCA categorization patterns.

3. **Speaker Gender Effect**: This is the novel prediction of this experiment: speaker gender manipulations will change the interpretation of the intermediate step C1 Fricative, and thus the direction of its PCCA influence on C2. For the likely perceived woman speaker, this intermediate step will be consistent with an "SH" interpretation, and lead to a decrease in "K" responses. For the likely perceived man speaker, this will be consistent with an "S" interpretation, and therefore increase "K" responses.

3.1.1.2 Eyetracking Data

Next, I consider the related predictions for the eyetracking data gathered during this experiment.

1. **C2 Effect**: Most fixations to velar targets when C2 is the velar endpoint ("K or G"); least when C2 is the alveolar endpoint ("T or D"); intermediate fixation amounts to velar target for intermediate C2 step. This would indicate that our C2 continuum synthesis was effective and illustrate that our eyetracking data recover meaningful behavioral signals.

2. **PCCA C1 Effect**: More fixations on velar targets ("K or G") when C1 is "S" or "L", less fixations on velar target when C1 is "SH" or "R", intermediate effect for intermediate C1 step. This corresponds to a replication of the traditional PCCA categorization effect.

3. **PCCA on-line Effect**: the effect of C1 on C2 fixations will be detectable during C1 (allowing 150ms to program an eye-movement). This would indicate that listeners are using C1 information to make inferences about the nature of C2 in real-time.

4. **Speaker Gender Effect**: for the *asta-ashka* class, the direction of the influence of the intermediate C1 step will depend on the speaker gender guise. For the likely perceived woman, this should pattern like "SH" while for the likely perceived man it should pattern like "S".

5. **Speaker Gender on-line Effect**: Given that the social cues are available at the beginning of the stimulus, we predict the influence of gender on C1 behavior to be present as soon as differences in C1 fixation behavior emerge.

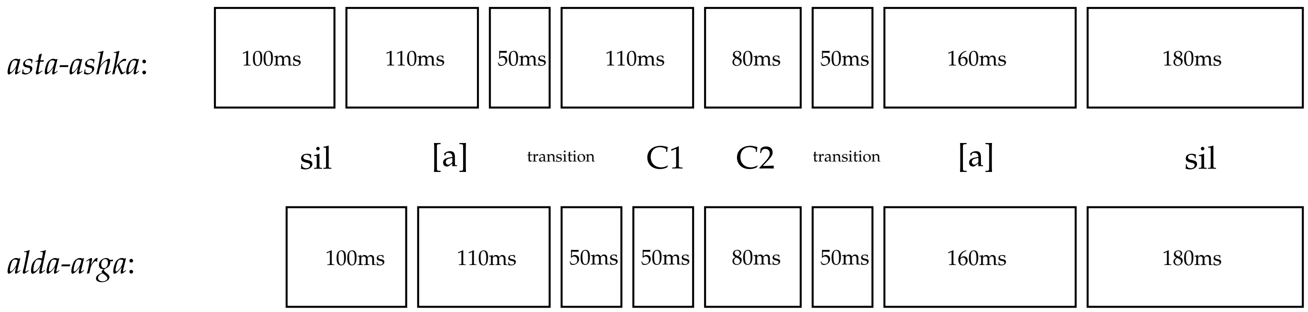


Figure 3.1: Duration of stimuli segments for both classes of stimuli. Width of boxes is proportional to segment duration. Gaps are included in this figure for visibility purposes, but stimuli segments are joined without gaps.

3.2 Methodology

3.2.1 Stimuli

Stimuli were created using an implementation of the Klatt speech synthesis system (Klatt, 1980) developed for Python by Ronald Sprouse and Keith Johnson.[1] Baseline synthesis parameter values were taken from recordings of various speakers producing the target items in initial-stress frames (e.g., [ˈaʃ.ta]). From these baseline recordings, endpoints were created for each individual stimuli segment. This involved manually smoothing formant and f0 trajectories, standardizing durations and intonational contours, and implementing more naturalistic stop bursts.[2]

The standardized durations for each stimuli class can be found in Figure 3.1. Tokens were synthesized at a sampling rate of 22050Hz. Because of constraints on sampling rate and KLP parameter refresh rate, the durations of the final stimuli items are approximately 12ms shorter than the requested durations shown in this figure (e.g., *asta-ashka*: 840ms requested vs. 828ms actual). Segment durations are identical across classes, except that the C1 durations for *alda-arga* are 60ms shorter than the C1 of *asta-ashka*, as initial impressionistic review by the author indicated that a shorter duration for these liquids was required to elicit a naturalistic percept. The speaker gender manipulation is largely carried out by changes in the vowel tokens. The intonational contours are held constant across speakers, with differences only manifesting in f0 and other formant parameters (F1-F5). For the *asta-ashka* class, the vowels (and their associated transition periods) are the only location of potential speaker differences: C1 and C2 are identical for all speakers. For the *alda-arga* class, speaker information is present during C1, as the acoustic realization of liquids required speaker-specific spectral information. After endpoints were established, inter-

[1] `https://github.com/rsprouse/klsyn`.

[2] Links to the KLP parameter files, stimuli recordings, and other files used in this experiment are available in Appendix A.

mediate steps were generated via a linear interpolation of KLP parameter values, setting the intermediate steps as exactly between the endpoint values.

Figures 3.2 and 3.3 present the waveforms and spectrograms for a subset of the endpoint syllables for the "likely perceived Woman" (W) and "likely perceived Man" (M) speakers. Note the presence of a brief transient at the beginning of the M *alda-arga* tokens. This is an artifact of the synthesis process and is low enough frequency to be inaudible.

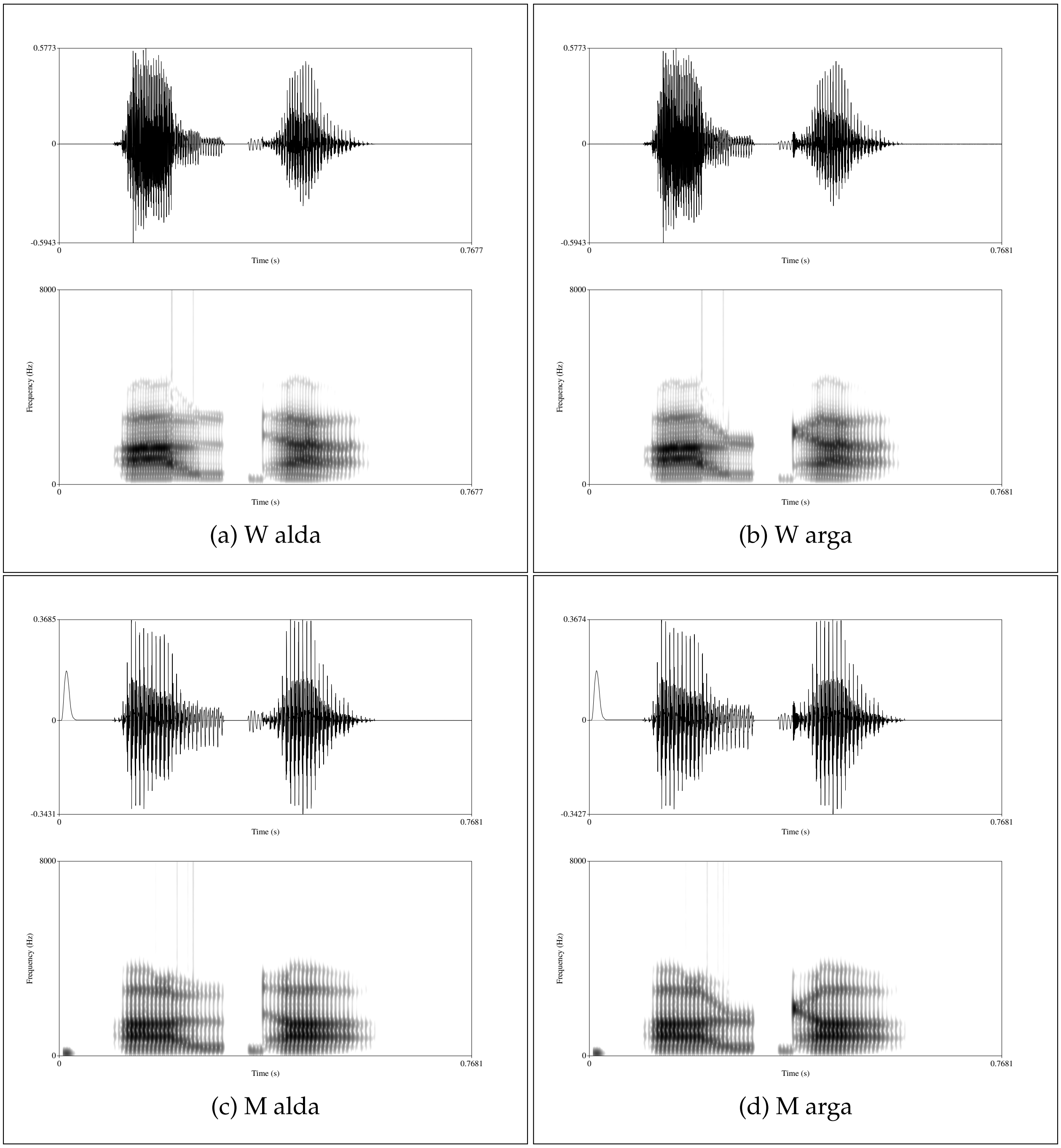
3.2.2 Participants

25 participants, largely Berkeley affiliated students and staff, took part in this experiment. Participants were recruited via flyers posted around campus, through email announcements, via snowball sampling, and from the Berkeley XLab participant pool. To be eligible for this study, participants were required to be 18 years of age or older, native English speakers (self-id), and have no known history of uncorrected speech, hearing, language, or vision impairments. As compensation for their participation, study participants received $10 cash. Recruitment, compensation procedures, and experimental methods were reviewed and approved by the UC Berkeley Committee for Protection of Human Subjects (#2021-11-14869).

Participants' ages ranged from 19 to 47 (mean: 21.48, median: 20). For the free response gender demographic question, 17 participants responded "female", 2 participants responded "F", and 5 participants responded "male". Given the uniformity in participants' responses to this free response question, I will split participants into two gender groups: female/F and male, named after the most common response for that group. As is typical of the subject population participants were drawn from, the demographic reported high levels of multilingualism, with nearly half of participants being simultaneous bi- or multilinguals. No participants were monolingual English speakers. However, the current experiment does not explore the role of multilingualism in this task. With such variability in participants' specific language backgrounds, I will not attempt to include this as a factor in these analyses.

3.2.3 Procedure

Before the experiment began, participants first completed the informed consent documents, the language background and demographic questionnaire, and received their compensation. The experiment itself took place in a WhisperRoom one-person sound booth, with participants seated in a chair approximately 20cm away from the desk. Stimuli were presented visually via a 300e 2nd Gen Lenovo Laptop with an 11.5" screen was placed 8.5cm away from the edge of the desk. The Tobii Pro Nano eyetracker was affixed directly below the center of the screen. Audio was presented through AKG 55 Ohms K240 Studio over-ear headphones and participants responded to experimental instructions via the laptop keyboard button presses and USB mouse clicks.



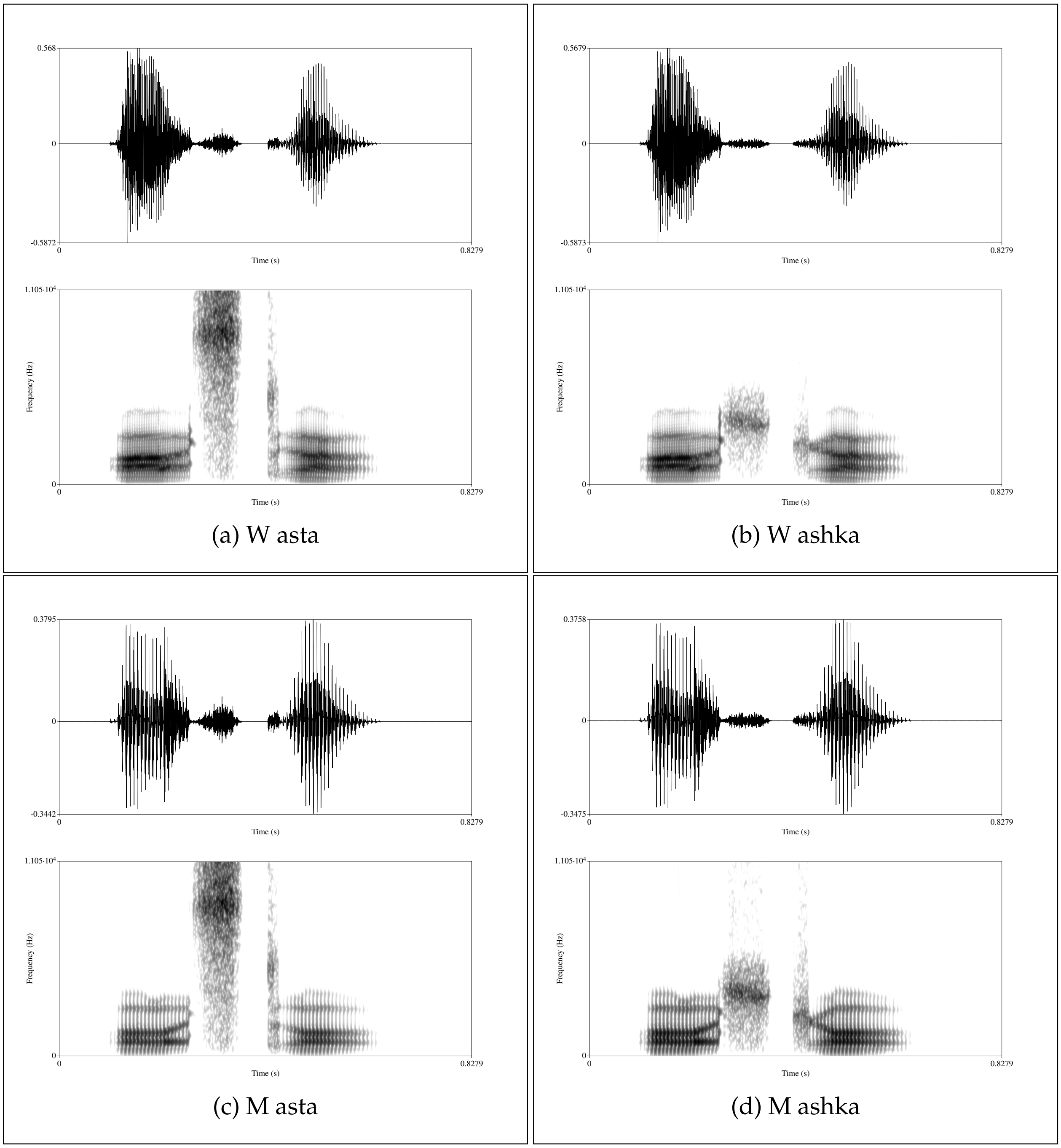
(a) W alda

(b) W arga

(c) M alda

(d) M arga

Figure 3.2: Waveforms and Spectrograms for endpoint *alda-arga* stimuli.



(a) W asta

(b) W ashka

(c) M asta

(d) M ashka

Figure 3.3: Waveforms and Spectrograms for endpoint *asta-ashka* stimuli.

The experiment was carried out within OpenSesame 3.3.10 (Mathôt et al., 2012) running on Windows 10 with a display resolution of 1366 x 768. The experiment consisted of three distinct phases: eyetracker calibration, an introduction to the task, and the main experimental trials. I utilized the default OpenSesame calibration routine provided by the PyGaze python package (Dalmaijer & Van der Stigchel, 2014). In this routine, participants fixated on points presented in different areas of the screen to estimate the participant's distance from screen, visual angle, and ensured that the detection algorithm's accuracy thresholds were met for that session.

In the next phase, participants were provided a brief introduction to the nature of the task through on screen prompts as well as a single practice trial. This practice trial was identical in structure to the main experiment trials (described below), except the stimulus item was a member of a non-test set ("AVA" - "ABA" - "APA" - "AWA") and it was presented in a separate resynthesized voice.

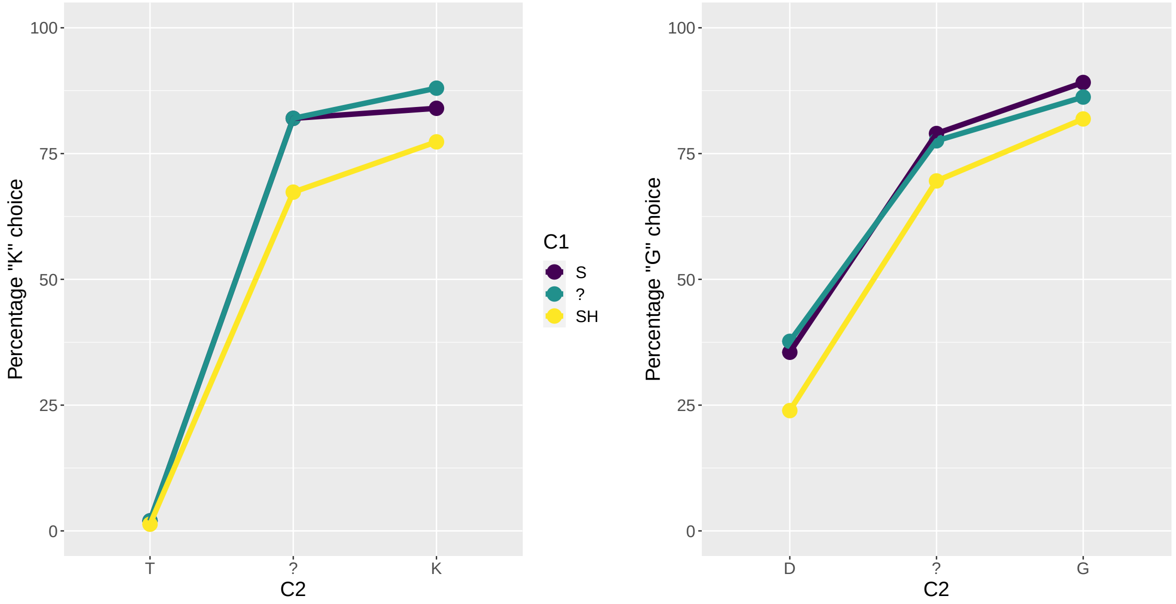
Then, participants moved to the main experiment block where they completed the 108 trials in a randomized order. The 108 trials are composed of 2 reps each of the 27 individual stimuli items for the 2 classes of stimuli (*asta-ashka* and *alda-arga*). In each trial, participants were first presented with a fixation cross at the center of the screen and were instructed to fixate on that point for its duration, 1000ms. Then, the orthographic stimuli options were presented at the center of each quadrant. Orthographic representations of each choice option were presented in the Upper Left, Upper Right, Lower Left, and Lower Right quadrants, centered on positions 192 pixels above or below, and 352 pixels to the left or the right of the center of the screen. The specific quadrant location of each orthographic choice was randomly assigned for a participant but remained fixed for them throughout the experiment.

Simultaneous with the end of the fixation cross and the presentation of the written choices, the trial audio began playing. Participants indicated their final choice via mouse click, at which point their cursor was automatically recentered on the screen and then the next trial began with its fixation cross presented at the center of the screen.

3.3 Categorization Results

3.3.1 Raw Data

During the course of the experiment, it was determined that the specific choice presentation arrangement for the first two participants was not correctly recorded. As such, the data for these participants indicated that participants made a selection in the Upper Left quadrant, for example, but did not explicitly correspond to that orthographic choice was presented in that quadrant. For the *asta-ashka* class, these two participants' responses to the unambiguous acoustic endpoints was nearly uniform, allowing us to infer with a high degree of confidence the presentation order for these trials. However, the *alda-arga* class exhibited greater variability in quadrant responses to each combination of stimuli



(a) *asta-ashka* class

(b) *alda-arga* class

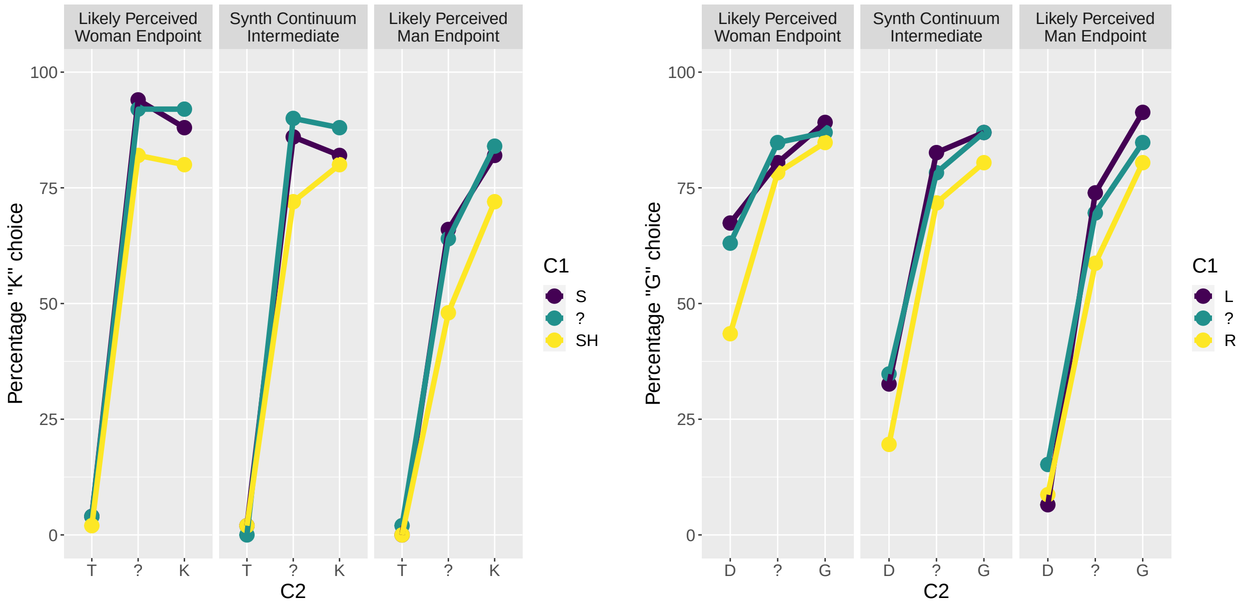
Figure 3.4: Categorization Aggregate Responses.

and the decision was made to exclude these trials from the subsequent categorization and fixation analyses.

This is observed within the aggregate mean categorization responses for each class, shown in Figure 3.4. Turning first to the *asta-ashka* items in Figure 3.4a, we can observe a clear separation between the "T" C2 items, which are nearly completely categorized as "T", and the intermediate step and "K" stimuli, which are mostly categorized as "K", regardless of condition. In these aggregate data, we do see evidence for the perceptual compensation for coarticulation (PCCA) effect in the expected duration for these two C2 steps: following a clear "SH", these two C2 stimuli are less likely to be categorized as "K". The PCCA effect is not visible in the "T" C2 step, since responses to this step of the continuum are effectively at floor.

Next, we consider the aggregate response data for the *alda-arga* class presented in Figure 3.4b. First, at the group level, responses to each of the C2 stimuli appear less categorical than the previous class, with "D" stimuli items not at floor. However, we still observe a clear PCCA effect in the predicted direction: "G" responses are more common following a clear "R" C1 for all C2 stimuli items.

Figure 3.5 splits the aggregate responses of Figure 3.4 by the speaker gender manipulation. Additionally, Figure 3.6 zooms in on aggregate responses to the intermediate "?" C2 stimuli, broken down by previous C1 and speaker gender manipulation. Going from left to right in each panel, we observe decreases in categorizations of C2 as the velar choice,



(a) *asta-ashka* class

(b) *alda-arga* class

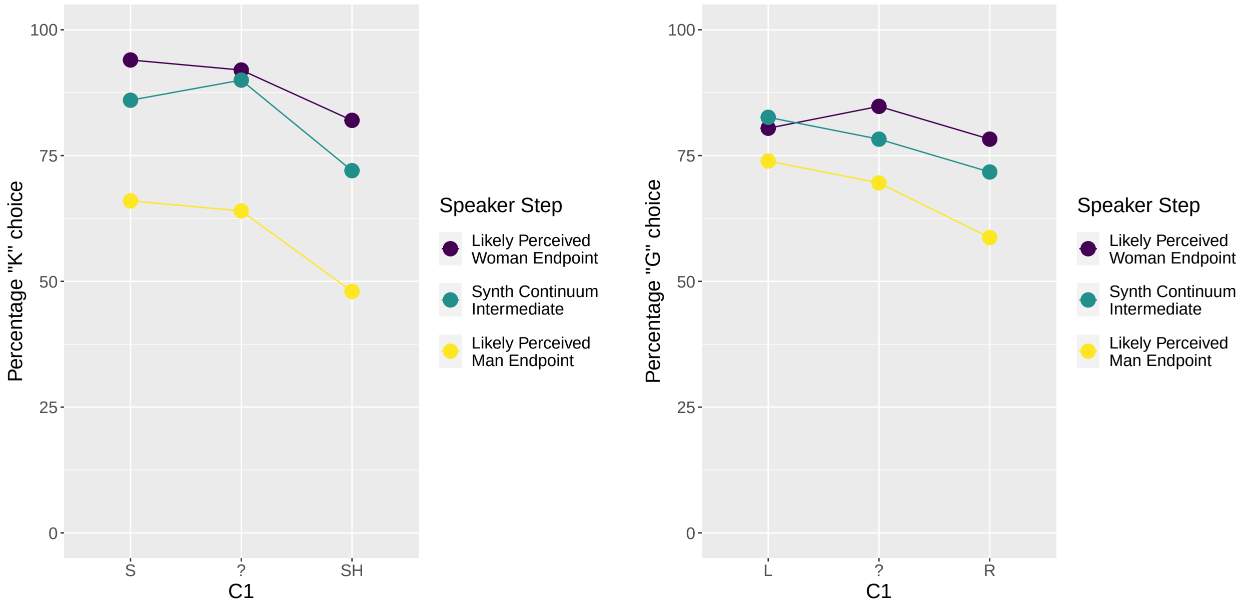
Figure 3.5: Categorization Aggregate Responses by Speaker.

consistent with the PCCA predictions. Additionally, for the *asta-ashka* class (and perhaps the *alda-arga* class as well), the entire C2 continuum seems less likely to be categorized as velar. While various suggestive differences are observed in the aggregated condition responses, I will reserve meaningful interpretation for the results of the statistical analyses below, which take into account not only differences in means but also variance in the data.

3.3.2 Model Results

Turning from the initial inspections of the raw data, I now consider the degree to which each of our predictions about the categorization data are statistically supported. Much like the models of the previous chapter, I draw upon Bayesian logistic models implemented in version 2.17.0 of the *brms* package (Bürkner, 2021). Again, arguments were kept to their default values, with the following exceptions: the response distribution family was bernoulli and the backend was CMDSTANR (v. 2.30.0).

Two series of models were fit to the categorization data: a series for the *asta-ashka* trials and a series for the *alda-arga* trials. The prior specifications, model-fitting procedures, and model outputs for each class will be discussed separately.



(a) *asta-ashka* class

(b) *alda-arga* class

Figure 3.6: Categorization Responses for intermediate step C2.

3.3.2.1 *asta-ashka* class

The first series of models were fit to the subset of the data containing only the *asta-ashka* class trials. Our dependent variable is the binary K_CHOICE which is 1 if the participant chose a k-word (“aska” or “ashka”) and 0 otherwise (“asta” or “ashta”). The models predictor variables are all ordinal variables: Syllable 1 Consonant Step (1 ‘likely ”S”’, 2 ‘intermediate step’, 3 ‘likely ”SH”’), Syllable 2 Consonant Step (1 ‘likely ”T”’, 2 ‘intermediate step’, 3 ‘likely ”K”’), and Speaker Step (1 ‘likely perceived Woman’, 2 ‘intermediate step’, 3 ‘likely perceived Man’). Given that each participant only heard 2 reps of each unique combination of these 3 variables, by-participant random effects are not appropriate.

All three independent ordinal variables are modeled as monotonic predictors. As further discussed for the Selective Adaptation models, this involves estimating both a simplex and scale parameter for each predictor. A weakly informative prior of $\mathcal{N}(0, 0.5)$ was chosen for the scale parameters, and the default uniform Dirichlet was chosen for the simplex parameters. As in the previous chapter’s model, an informative prior of $\mathcal{N}(-3, 1)$ was chosen for the intercept term. This represents our high degree of confidence that the combination of variables corresponding to our reference levels (‘likely “SH”’, ‘likely “T”, “likely perceived Woman”) would elicit the least “K” responses.

To determine the best fitting model, I fit the nested models presented in Table 3.2, corresponding to no-interactions, all pairwise interactions, and a complete interaction between all 3 predictor variables.

m0	C1 + C2 + Speaker Step
m1	C1 * C2 + Speaker Step
m2	C1 + C2 * Speaker Step
m3	C2 + C1 * Speaker Step
m4	C1 * C2 * Speaker Step

Table 3.2: Nested model interaction structure.

	elpd_diff	**se_diff**	**elpd_loo**	**se_elpd_loo**	**p_loo**	**se_p_loo**	**looic**	**se_looic**
m2	0.000	0.000	-463.267	21.302	6.792	0.545	926.535	42.605
m0	-2.294	1.841	-465.561	21.186	5.586	0.398	931.122	42.372
m4	-2.789	0.869	-466.057	21.522	10.273	0.835	932.113	43.045
m1	-3.094	2.027	-466.362	21.377	7.259	0.592	932.723	42.754
m3	-3.497	1.899	-466.765	21.290	7.176	0.489	933.530	42.579

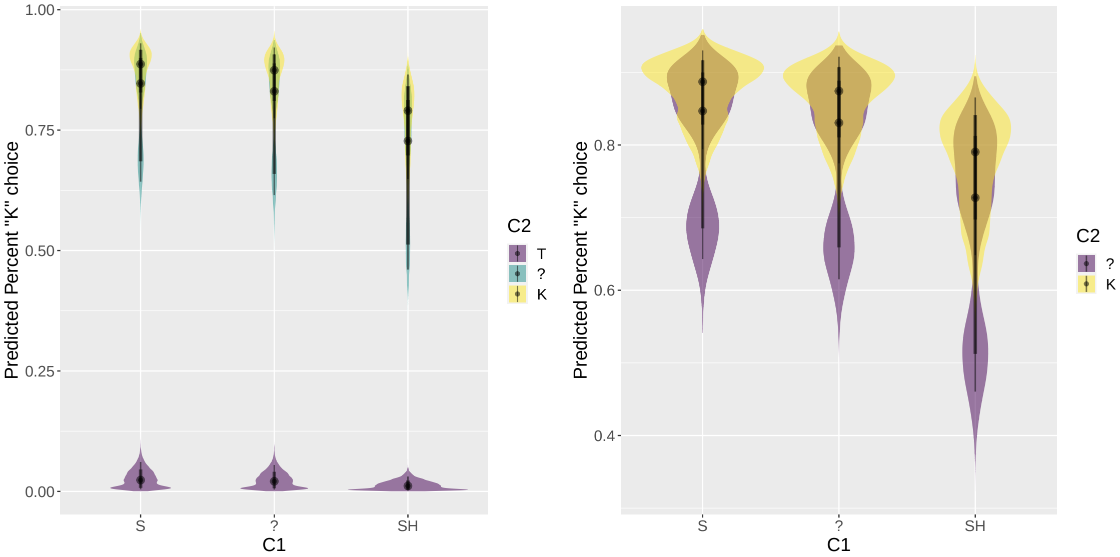
Table 3.3: *asta-ashka* model fitness information criteria comparison.[3]

Model fitness was determined by comparing various measures of models' LOO (leave-one-out cross-validation) information criteria, as implemented by the LOO_COMPARE function in BRMS. Like more familiar information criteria such as AIC or BIC, LOO estimates should not be interpreted directly, but rather in relation to the LOO estimates of a competitor model. LOO estimates are calculated via the leave-one-out cross-validation process. The various LOO estimates and their standard errors are presented in Table 3.3. m2 (an interaction between C2 and Speaker Step) is the best fitting model (lowest looic, closest elpd to 0), but is only a slight improvement over the no-interaction model (m0). This indicates that we have weak evidence for the Syll. 2 and Speaker interaction, and no evidence for any remaining interactions.

Let us now consider the best-fitting model, m2. This model includes an interaction between Speaker Step and C2, but no interactions with C1. Given that interpreting the posterior parameter estimates for multiple ordinal predictors in a logistic regression can be unintuitive, I instead present random draws from the fitted posterior distribution of all parameters, allowing us to more easily map onto probability estimates.

Observing the effect of C2 on the posterior distribution in Figure 3.7, our Prediction 1 as well as the patterns observed in the aggregate responses are well supported: the "T" stimuli are consistently at floor, with very narrow distributions. The other C2 steps

[3] elpd_loo: expected log pointwise predictive density; elpd_diff: difference in elpd_loo between current and best-fitting model; p_loo: effective number of parameters; looic: -2*elpd_loo



(a) All C2 Steps

(b) Only Intermediate and "K" C2 Steps

Figure 3.7: Posterior distribution draws by C1 and C2 Step.

overlap one another and have wider distributions, as made especially clear in the zoomed in version of Figure 3.7b. The distributions of the intermediate "?" step tend towards "K" despite being the intermediate continuum step, and are especially wide, exhibiting clear evidence of a bimodal distribution.

This bimodal distribution corresponds to the differential influence of speaker gender manipulations on ratings of C2. Figure 3.8 represents the posterior distribution draws for only the intermediate C2 step. Recall that across gender conditions, the C1 acoustics are kept constant. That is, the "SH" step presented along with the "likely perceived man endpoint" voice is identical to the "SH" step presented with the other two speakers. When speakers are separated, the PCCA effect is clearly obtained: for all speakers, the posterior distribution of "SH" is further to the left than the other C1 steps. We can thus confirm with a high degree of certainty our Prediction 2: the intermediate C2 step is less likely to be categorized as "K" following "SH".

Regarding our Prediction 3, we do not observe the differential effects of perceived gender on the PCCA inducing behavior of C1 on C2. Recall that our prediction was that the intermediate C1 would be variably perceived for the different speakers (more likely perceived as "S" for the "likely perceived man endpoint", and more likely perceived as "SH" for the "likely perceived woman endpoint"). However, a model which included an interaction between speaker gender step and C1 did not significantly improve model fitness, leading to us to conclude that the effects of C1 are consistent across all speaker gender manipulations. What we do observe, however, is an overall difference in C2 responses to each of the speaker gender guises. This is clearly observable in Figure 3.6a in which we

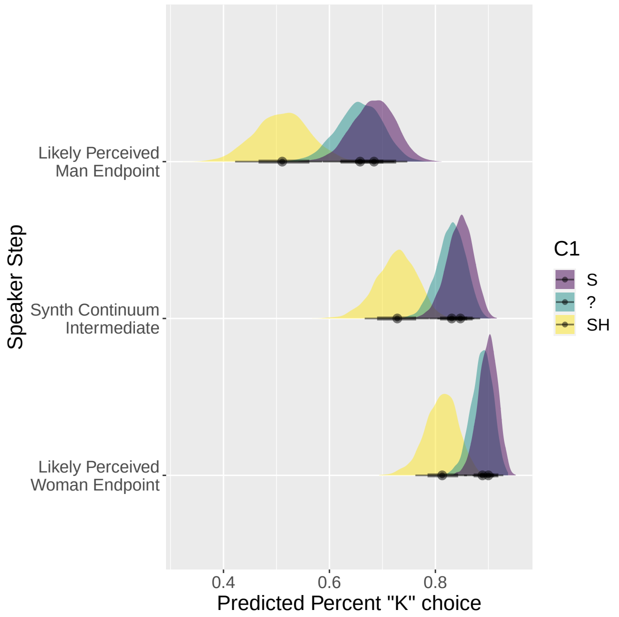


Figure 3.8: Posterior distribution draws for intermediate C2 step by C1 step and speaker step.

observe clear evidence for the “likely perceived man endpoint” speaker eliciting less “K” responses and more “T” responses overall, compared to the other two speakers. This effect was not included in our initial predictions, but is quite robust. I return to this question in Section 3.3.3.

3.3.2.2 *alda-arga* class

The model fitting process for *alda-arga* class proceeded identically to the previous class. Like the *asta-ashka* trials, the best fitting model for this class was m2, which included an interaction between C2 and speaker gender step, but no interactions with C1. This lack of support for our Prediction 3 is consistent with the analyses of the *asta-ashka* data.

Observing the draws from the posterior distribution by C1 and C2 step in Figure 3.9, it is immediately obvious that this class of stimuli exhibit a higher degree of variance in responses than the previous class. While the intermediate and “G” steps are relatively well separated and narrow, the posterior distributions for the ‘likely “D” endpoint’ are extremely wide. This correlates to a measurable difference in the categorization of this C2 step dependent upon the speaker gender manipulation, as presented in Figure 3.10. While the likely “D” step is categorized nearly completely as “D” for the “likely perceived man endpoint” speaker, this step is often categorized as “G” when presented in the other voice gender guises.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
m2	0.000	0.000	-621.829	19.217	8.769	0.397	1243.658	38.433
m4	-2.641	0.561	-624.470	19.376	11.894	0.572	1248.940	38.752
m0	-12.854	4.740	-634.683	19.121	6.484	0.270	1269.366	38.241
m1	-13.742	4.751	-635.571	19.189	7.408	0.321	1271.143	38.378
m3	-13.836	4.744	-635.665	19.171	7.506	0.303	1271.330	38.343

Table 3.4: *alda-arga* model fitness information criteria comparison.

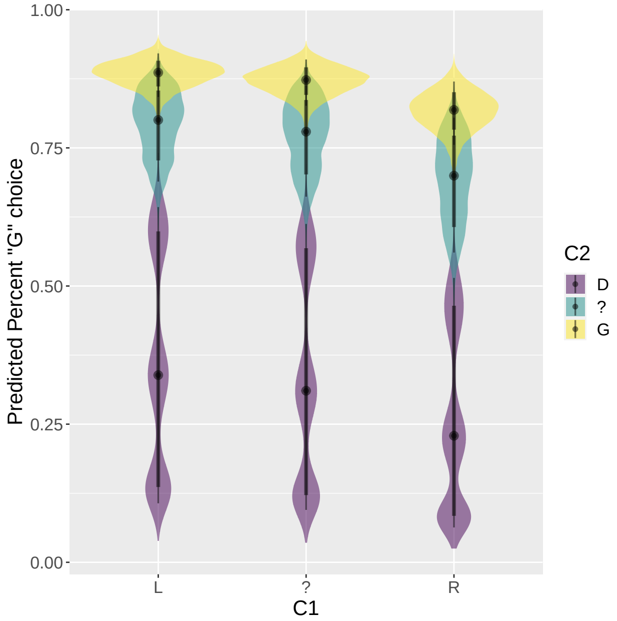


Figure 3.9: Posterior distribution draws by C1 and C2 step.

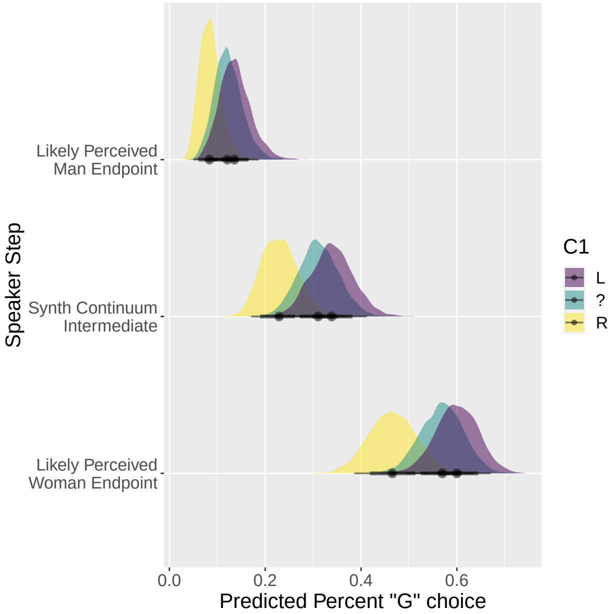


Figure 3.10: Posterior distribution draws for 'likely "D"' C2 step by C1 and speaker step.

Despite this variability, we do observe evidence for both our Prediction 1 (main effect of C2 step) and our Prediction 2 (more "G" responses after "L" than "R"). The overall PCCA effect is especially evident in the case of the intermediate step C2 shown in Figure 3.11. As is consistent with our predictions and previous PCCA literature, we find clear evidence for a decrease in "G" categorizations following a clear "R" C1 than the intermediate or "L" C1 steps.

As discussed previously, the lack of statistical support for an interaction between C1 and speaker gender leads to us not finding evidence supporting our Prediction 3. Like the *asta-ashka* class, we do however see in Figure 3.11 an overall decrease in "G" categorizations for the "likely perceived man endpoint" speaker than the other two speaker gender steps. As in the previous class, this effect was not predicted, and I return to this in the following section.

3.3.3 Discussion

To summarize, the categorization data from this experiment replicated two well-known instances of perceptual compensation for coarticulation (PCCA). Consistent with this previous work and our predictions, perception of a preceding fricative (in the case of *asta-ashka*) or liquid (in the case of *alda-arga*) significantly shifted listeners' categorization of the following stop's place of articulation.

Regarding our novel Prediction 3, we did not observe evidence for a significant influ-

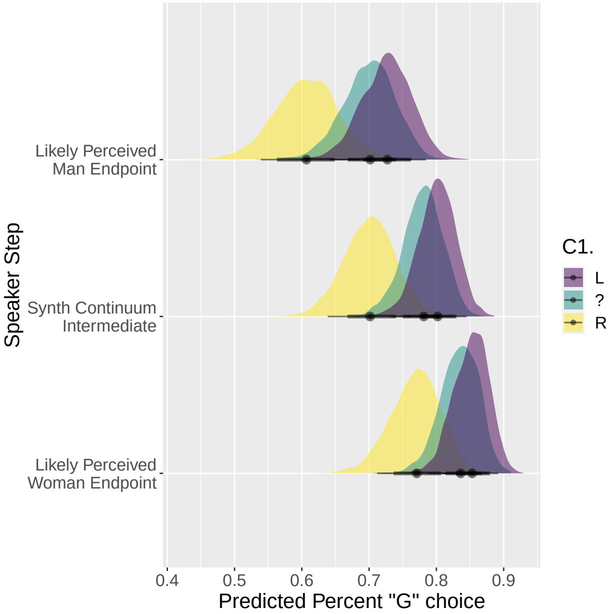


Figure 3.11: Posterior distribution draws for intermediate C2 step by C1 and speaker step.

ence of speaker gender guise on the PCCA behavior of intermediate C1 fricatives. Recall that if an intermediate fricative would conceivably be treated as an /s/ when presented in a man's voice, but /ʃ/ in a woman's voice, we would predict this token's PCCA effects to shift categorization of the following C2 in opposite directions. However, this prediction rests on the assumption that the intermediate fricative stimulus is plausibly interpreted as different phonemes in the different speaker guises. From the raw categorization data as well as the modeling results, it is evident that our intermediate C1 steps were not perceptually intermediate, and were effectively treated as "S" and "L" for all listeners. As such, we do not have the appropriate level of uncertainty or intermediateness in the underlying phoneme identity of the fricatives to be able to adequately test our Prediction 3.

We also observe an unpredicted effect of the speaker gender manipulation on baseline C2 response rates. That is, in both classes of stimuli the categorization of C2 for likely perceived man speaker was shifted away from the velar C2 category and towards the alveolar C2 categorization. This is most visible for the intermediate C2 posterior draws for *asta-ashka* shown in Figure 3.8, where a categorization of "K" for this speaker is nearly 30% less likely. This could be due to differences between the speakers in magnitude of formant transitions following the vowel. For all speakers, the duration of formant transitions following C2 into the stable portion of the final [a] was kept constant. However, the exact F2 and F3 values during the stable portion of the vowel vary between speakers, leading to slight differences in the magnitude (or "steepness") of the formant transitions. This could potentially lead to different interpretations of the place of the previous burst, and

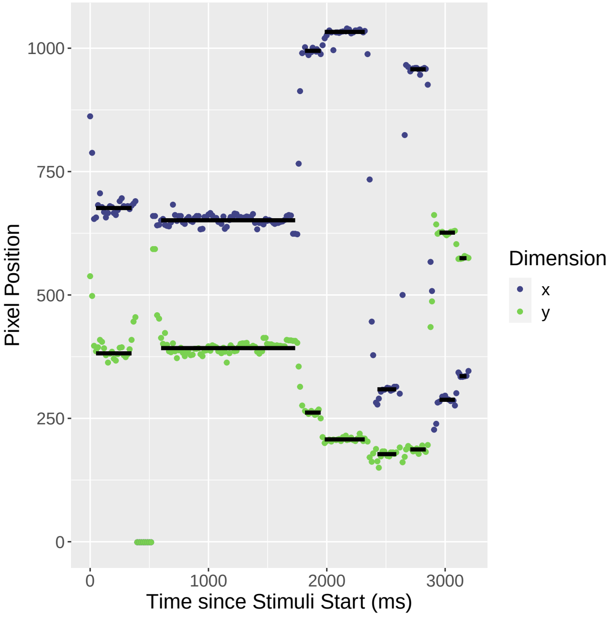


Figure 3.12: Example of SACCADES output from a single trial. Points represent individual measurements, lines indicate automatically detected fixations.

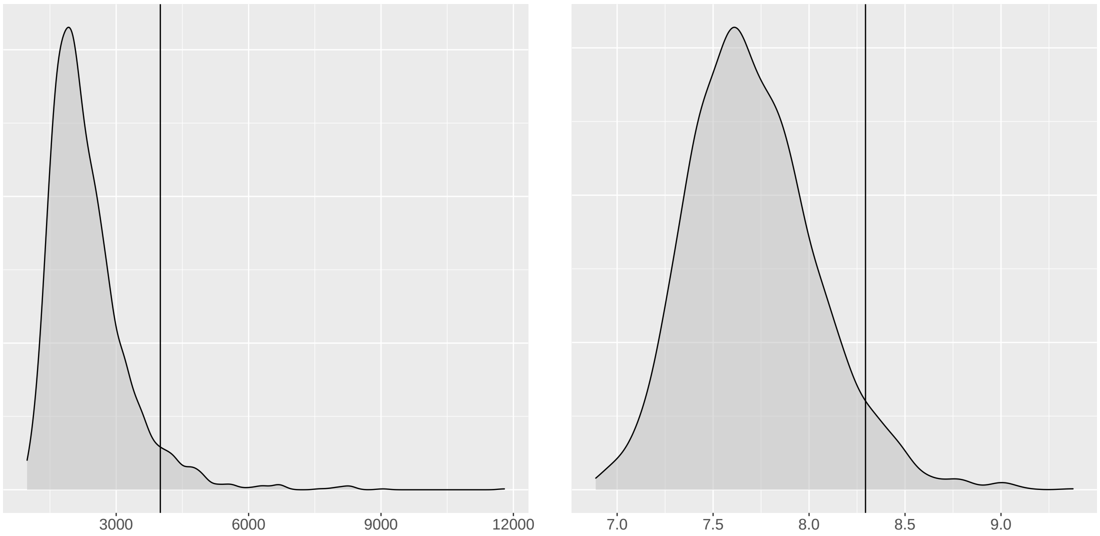
cause the C2 categorization difference by speaker gender that we observe here.

3.4 Eyetracking Results

3.4.1 Data Pre-Processing

The first step in data processing involved transforming the coordinate system of OpenSesame's mouse clicks and cursor movement (center of screen is the origin) to the coordinate system of the Tobii gaze measurements (upper left corner of screen is the origin). Next, the SACCADES R package (v. 0.2.1, von der Malsburg, 2019) was used to automatically detect events in the raw gaze position data. This detection algorithm relies on velocity measurements to detect periods of rapid transition and movement (saccades) and periods of relative stability (fixations or blinks). Given the relative low sampling rate of the eyetracker used, I followed the procedure outlined by the SACCADES author and enabled coordinate smoothing and disabled saccade smoothing, though qualitatively this did not appear to significantly affect the number, duration, or location of fixations detected.

An example application of this method to an individual trial is presented in Figure 3.12. The automatically detected fixations are noted by the black lines. This trial includes a clear example of a blink at approximately 500ms, with a characteristic shift in the y-axis estimates in the lead up to and transition out of the blink (pixel position estimates are 0).



(a) Response time - ms

(b) Response time - log(ms)

Figure 3.13: Time from start of stimuli to response. Vertical line represents the response cut off of 4000ms.

Smoother periods of rapid transition in both x and y axes are characteristic of saccades, as evidence by the saccade between transitions in this trial at approximately 1750ms.

Following the automatic fixation detection algorithm, I excluded fixations that occurred within 60 pixels of a quadrant boundary. This threshold was determined by visually inspecting the distribution of fixations across the experiment and excluded 34% of fixations from consideration. These 34% excluded fixations include fixations on the center fixation cross (which by definition occurs at the start of every trial), as well as fixations that were not located near enough to the target choices. Finally, adjacent fixations which occurred within the same quadrant but that occurred at slightly different positions within that quadrant were treated as a single fixation for the purposes of the quadrant-based analyses below.

I took a data-driven approach to determining what the maximum response time for analysis should be. Observing the distribution of response times in Figure 3.13, we can observe that these are well behaved and appear log-normally distributed. 95.2% of trials were completed faster than 4,000ms, and I take this as my upper bound, excluding any trials with a longer response time.

3.4.1.1 *asta-ashka* class

3.4.1.2 Raw Data

Let us first ensure that the fixation data are sound before exploring the effects of our experimental manipulations. Figure 3.14 presents the percentage of fixations in each 100ms bin[4] as calculated from the raw fixation counts. Each panel represents trials where participants made a particular choice. For example, the top left panel represents fixations in trials where listeners ultimately chose "ashka". Unsurprisingly, we see a peak in "ashka" fixations in this panel, as represented by the dark purple line representing these fixations. In all four panels, the target that is eventually chosen receives the most fixations, serving as an indicator that the fixation data are sound and reflect participants' eventual decision.

While our earlier categorization analyses indicate that listeners' classification behaviors are linked to our experimental manipulations, the relationship is not perfect. As such, I turn away now from an investigation of fixation trajectories by listeners' ultimate choices, and instead explore how fixations vary as a function of our experimental manipulation of C1, C2, and speaker gender.

Figure 3.15 presents the raw fixation trajectories by C1 and C2 condition averaging across speakers. For each condition I am plotting a separate trajectory for each potential fixation target. This view allows us to investigate how listeners' fixation patterns evolve during the course of the trial. For example, consider the top right panel which corresponds to the clear "asta" condition: in this panel, we see that fixations on "asta" and "aska" quickly rise in tandem over the "SH" targets, before the "asta" competitor falls off at approximately 700ms, with "aska" receiving the bulk of later fixations.

The patterns we observe here tend to echo what was found in the classification trials: the endpoint C1 and C2 manipulations elicited perceptual behavior as predicted but the intermediate steps chosen for this experiment do not appear to be perceptually intermediate. Returning to Figure 3.15, we can see that the intermediate C1 fixations closely mirror that of "S", while the intermediate C2 fixations closely match those of "K". Given that our intermediate steps appear (averaging over speakers) to not be truly perceptually intermediate, this may prevent us from testing for fine influences of speaker gender that we predicted would be strongest for the most ambiguous consonant steps. I return to this point in the later discussion section.

In addition to the raw fixation trajectories, another useful measure of the influence of our experimental manipulations on C2 fixation behavior is the calculation of a Fixation Bias measure. Following similar methods utilized in McMurray et al. (2008) and Galle et al. (2019), I first calculate for each bin the percentage of fixations to "K" targets and to "T" targets. Then, I calculate $Bias_{k-t}$ by taking the difference of these two percentages.

[4] In actuality, this includes an implicit 5th category of "no fixation". For example, in the 0-100 bin nearly no one is fixating in one of the quadrants, and the sum of the percentage fixations of the four targets in this bin is quite small. If the "none/not fixating" were not included, the legibility and informativity of this plot would decrease, with areas with low fixation counts varying wildly.

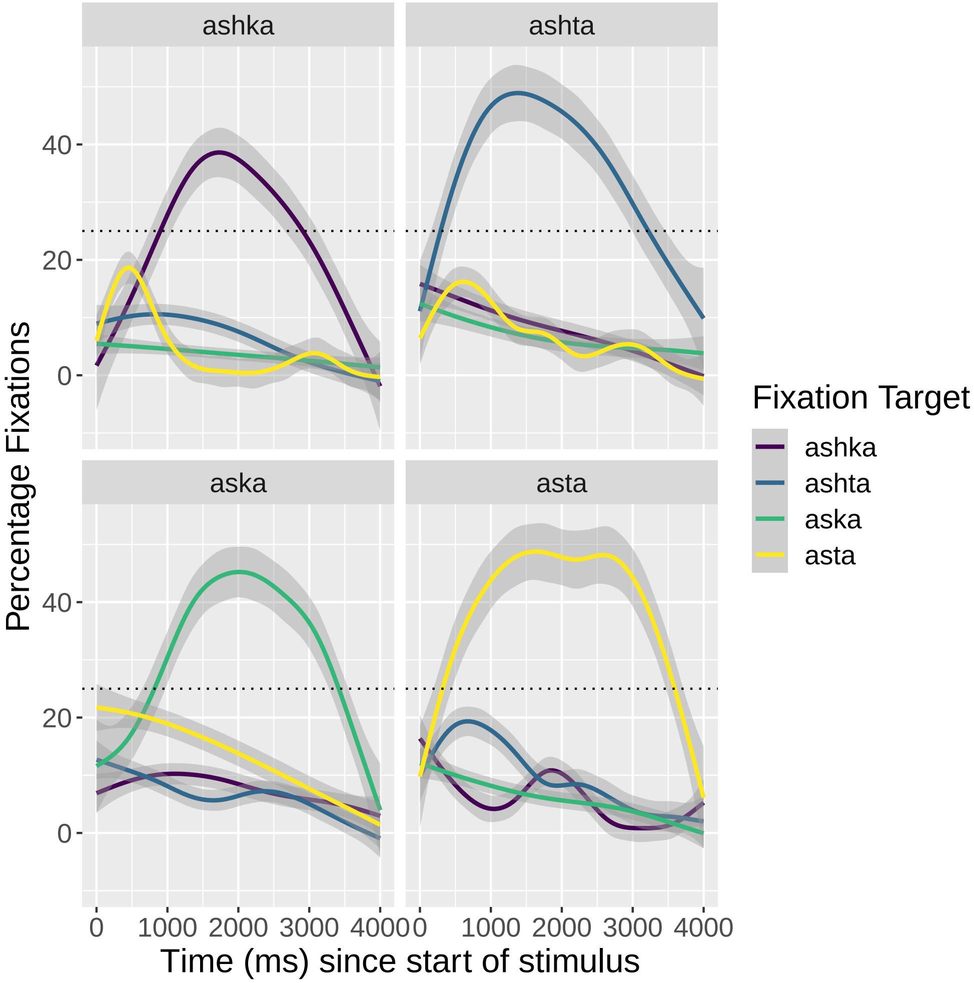


Figure 3.14: Percentage of fixations on specific *asta-ashka* Fixation Targets, broken into panels representing listeners' final decisions.

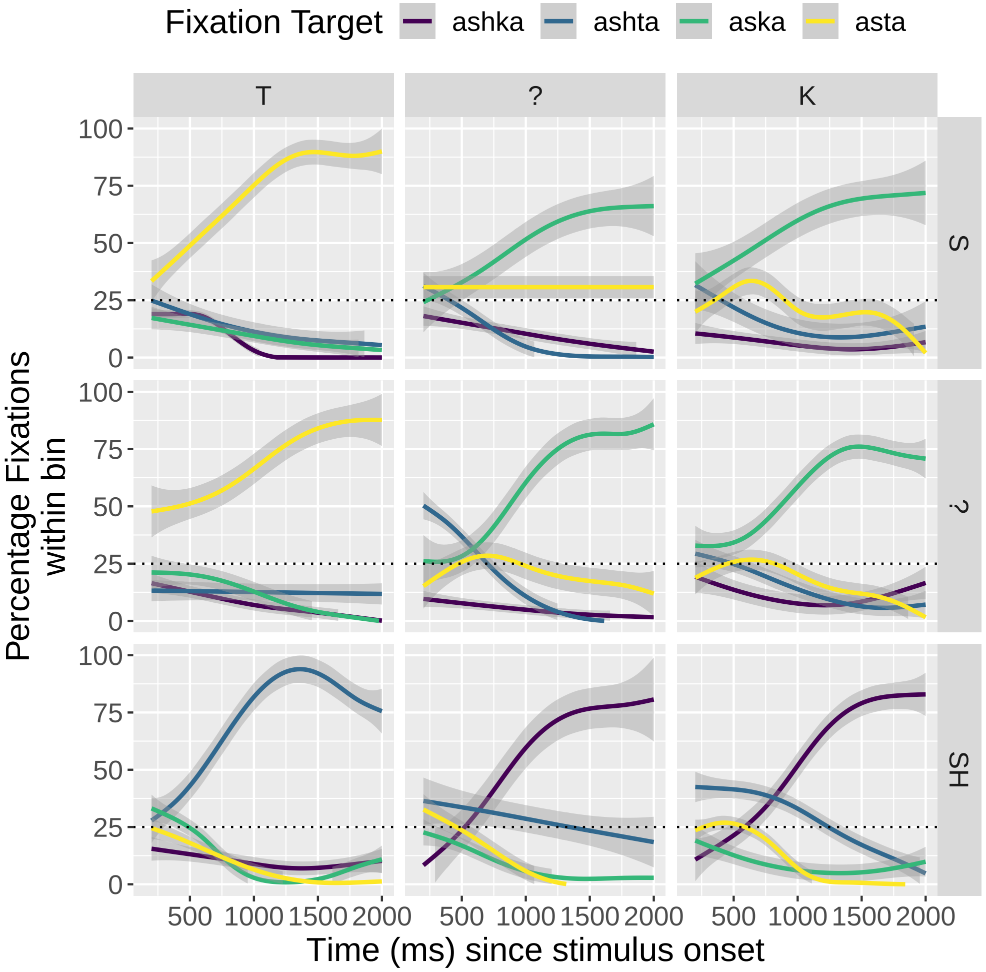


Figure 3.15: Smoothed fixation percentage by C1 and C2.

This new measure can range from 1 (in this bin everyone was fixating and they were fixating on “K”) to -1 (in this bin everyone was fixating and they were fixating on “T”). Note that a $Bias_{k-t}$ value of 0 could map onto a situation in which there are no fixations (0% “K” - 0% “T” = $Bias_{k-t}$ of 0) all the way up to a scenario in which everyone is fixating (50% “K” - 50% “T” = $Bias_{k-t}$ of 0); in any case, the percentage of fixations to “K” and “T” are equal. This reflects the fact that we are not primarily concerned with how many fixations occur in any given bin, but rather whether each bin is biased towards a given C2 fixation target.

Figure 3.16 presents the $Bias_{k-t}$ values, averaging across conditions and broken down by C2 step. We can observe that the $Bias_{k-t}$ values for all three C2 steps are all overlapping initially, and are slightly negative, indicating an overall bias towards fixating on “T” in this period, regardless of condition. As we will see later, this does not represent the

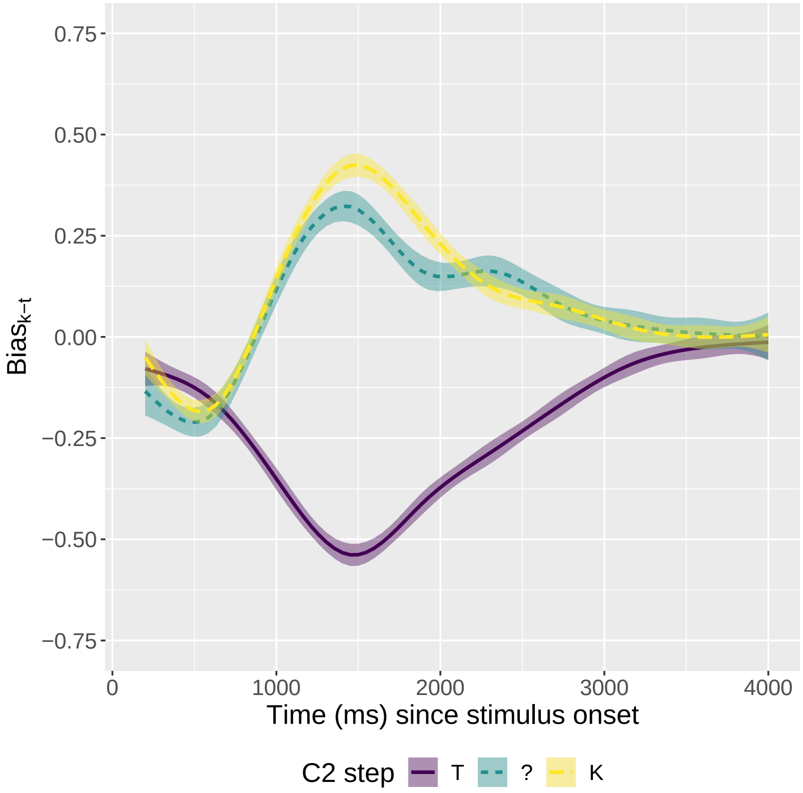
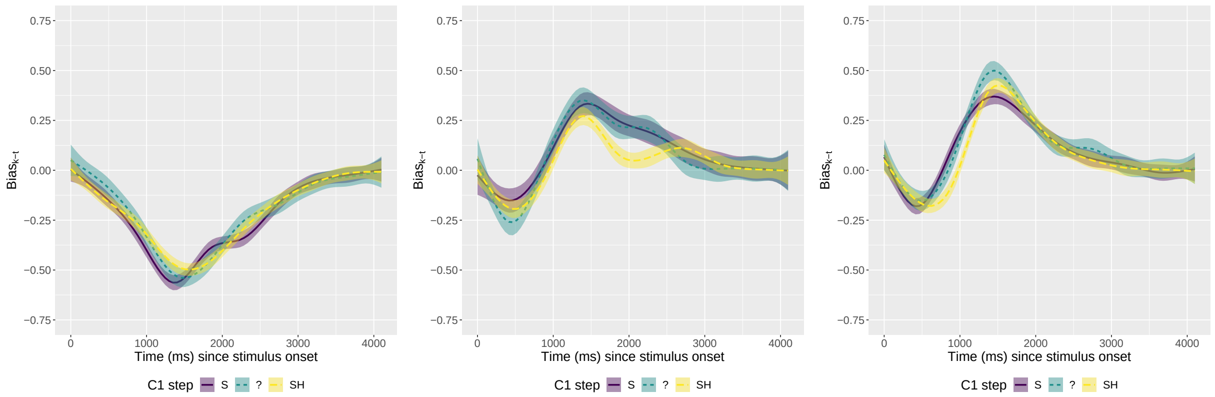


Figure 3.16: $Bias_{k-t}$ averaging across conditions.

hidden effect of an experimental manipulation, as this effect is present in each individual condition as well. This pattern of early fixation bias to "T" also does not represent a confound based on any visual preferences for specific quadrant positions, as the quadrant location of each fixation target was randomized between speakers. As such, the likeliest interpretation is that this difference represents an overall early bias towards "T" targets, regardless of the specific conditions.

Moving past the early bias towards "T" fixations, we see the detectable emergence of the C2 influence on fixations occur starting around 700ms, when the "T" and "?/K" stimuli begin to diverge, with the latter peaking at approximately 1500ms with a positive value. This indicates that this period exhibits the greatest $Bias_{k-t}$ and thus the greatest difference between percentage fixations to "K" and fixations to "T". Interestingly, the intermediate step C2 "?" induces a bias that is temporally aligned with the endpoint "K", but slightly weaker. This indicates that we do not have evidence for any additional processing time associated with an ambiguous token, and most listeners perceived this step equivalently to "K".

I turn now to the question of potential PCCA effects induced by differences in C1 step,



(a) C2 - T

(b) C2 - ?

(c) C2 - K

Figure 3.17: $Bias_{k-t}$ separated by C1 and C2 steps.

and examine in Figure 3.17 the bias measures calculated separately for each C1 and C2 combination. Recall that the PCCA prediction is that we expect less fixations to "K" (a smaller $Bias_{k-t}$ value) following "SH" than following "S". The patterns presented here are suggestive, and we see evidence for the predicted difference in fixations, particularly in the central intermediate C2 panel. However, it remains to be see whether these patterns persist when we not only consider the influence of speaker gender, but also when subject our interpretations and predictions to statistical analysis.

3.4.1.3 Modeling

To determine whether the patterns observed in these initial explorations of the raw data represent stable and meaningful trends, I now turn to statistical modeling of the *asta-ashka* subset of the data. Given the time-varying nature of the signal, I will utilize generalized additive models (GAMs) during the modeling phase. This modeling approach allows for robust investigations of variable time-course data, without specifying *a priori* the shape of the response distribution or factor smooths (e.g., linear, cubic, etc.). Like the categorization models, our fixed effects are C1 step, C2 step, and Speaker Gender step; we do not have enough observations per participant in each condition to include participant as a random effect.

Our dependent measure is $Bias_{k-t}$ as calculated in the previous section; this variable represents the difference of percentage "K" fixations and percentage "T" fixations for each bin. A value of 0 would indicate that participants are fixating on "T" and "K" equally during that bin. A value of -1 in a given bin would indicate that participants were only fixating on "T" in the bin, with no fixations on "K".

GAMs were fit using the MGCV R package (Wood, 2011) with various visualizations constructed using functionality of the ITSADUG R package (van Rij et al., 2022). All mod-

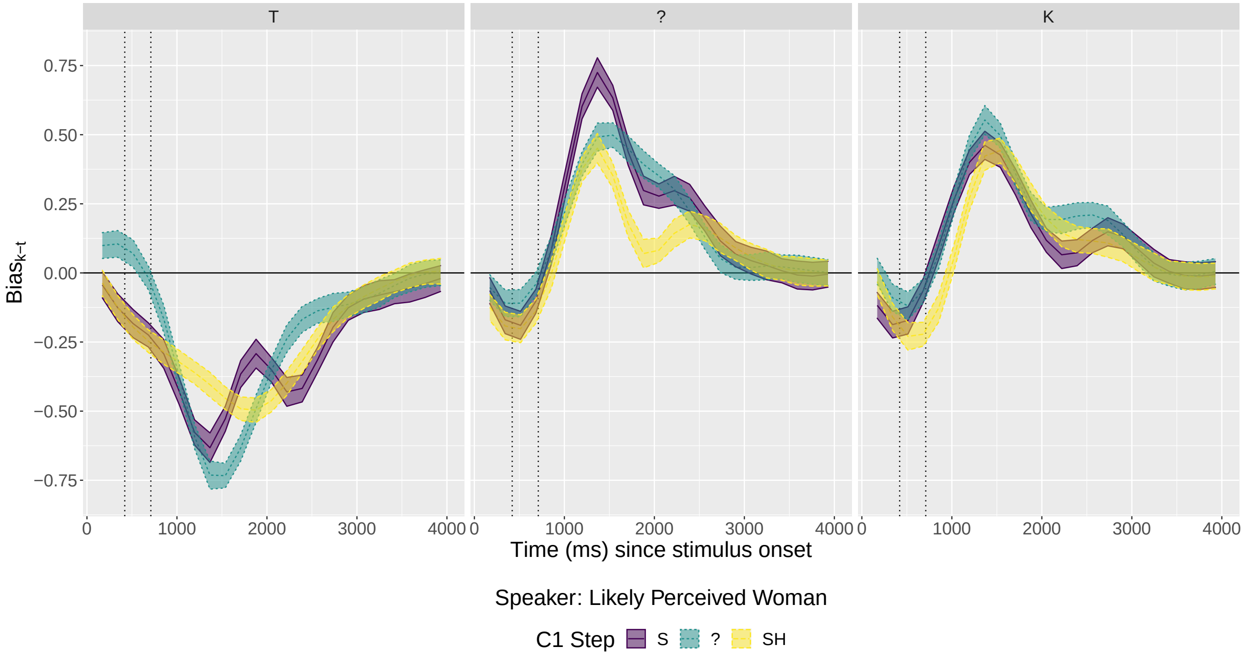


Figure 3.18: Model predicted $Bias_{k-t}$ for the likely perceived woman speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

els included smooths of bin start time using cubic regression splines as basis functions. I compared all candidate models created by the inclusion of the 3 fixed effects, their interactions, and their influence on the bin start time smooth (using the "by" functionality). Model fitness was determined via AIC, with the best fitting model being one including a smooth of time by the three-way interaction of our fixed effects.

Figure 3.18 presents the model results for only the "likely perceived woman" speaker. For these and all related visualizations of model predictions, non-overlap between two conditions (or a condition and a value, say 0) indicates a significant difference at that specific time point.[5] Predicted $Bias_{k-t}$ is plotted on the y-axis, and we can observe that our C2 manipulations significantly condition participants' fixation strategies. The leftmost panel presents fixations during trials where the C2 step was "T"; unsurprisingly, $Bias_{k-t}$ is predicted to be largely negative, indicating a strong bias towards fixating on "T" rather than "K" in this trial. Interestingly, all three C2 panels demonstrate a slight but measurable bias towards early fixations to "T" in almost all conditions, as seen in the dip below 0 from approximately 150-600ms. This is a true bias towards "T" "first guesses", rather than simply a positional quadrant bias, as the specific quadrant position of targets were

[5] Though, be sure to consider the granularity of the data and the limitations therein. Bias is calculated for every 100ms bin and the eyetracking samples occur approximately every 17ms.

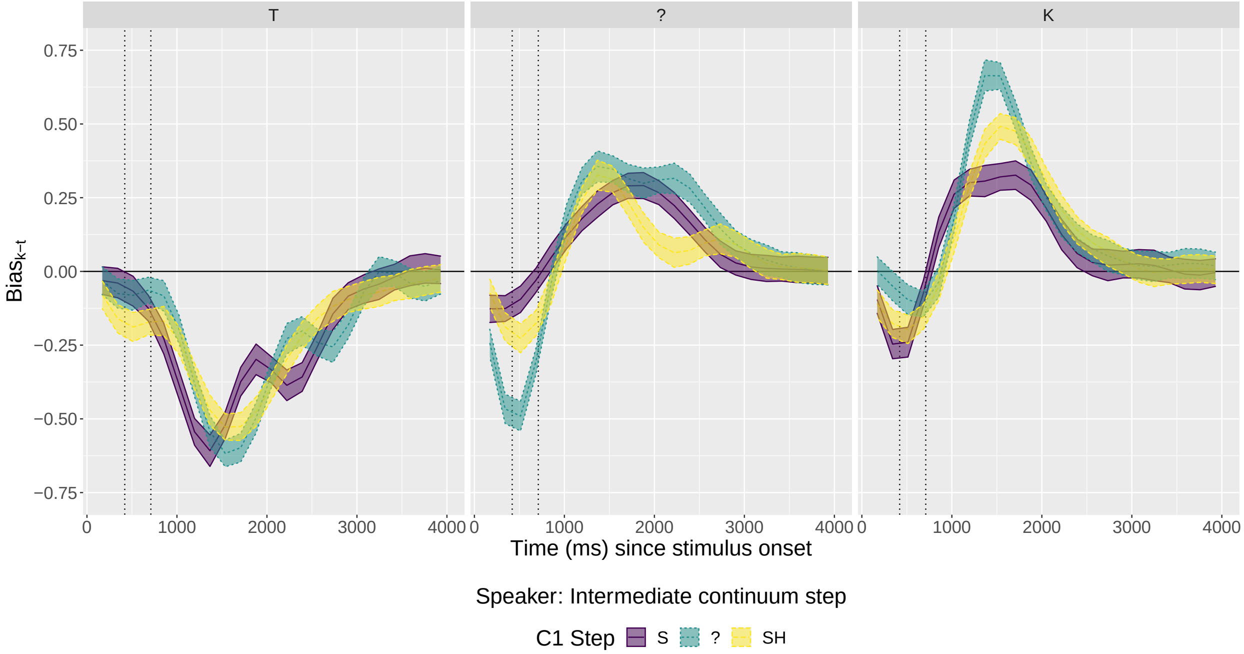


Figure 3.19: Model predicted $Bias_{k-t}$ for the intermediate step speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

randomized between speakers.

These patterns continue for the remaining two speakers in Figures 3.19 and 3.20, with evidence in both of an early "first guess" to "T", regardless of condition. Additionally, we continue to find clear support for our Prediction 1 for these speakers, with as well as clear effects of the C2 manipulation, supporting our Prediction 1.

I now turn towards testing our Predictions 2 and 3 which involve the PCCA effect of C1 on C2 fixations. First, in Prediction 2 I outlined the expectation that, like the classification data, we should see greater fixations to "K" following a clear "S" than after "SH". Additionally, we would expect this PCCA effect to be most visible when C2 at its most ambiguous, at its intermediate step.

Figure 3.21 presents precisely this comparison from the model. The y axis now represents the predicted difference in $Bias_{k-t}$ for "S" C1 trials versus "SH" C1 trials for only the intermediate C2 step. A positive value indicates that the $Bias_{k-t}$ effect is larger following "S" (more bias towards "K"), while a negative value, the opposite (more bias towards "K" following "SH"). A value of zero indicates no detectable differences in $Bias_{k-t}$ between the two C1 conditions.

We find evidence for a greater bias towards fixating to "K" following "S" for both speakers, as indicated by the portions of the trajectories significantly above 0 which are

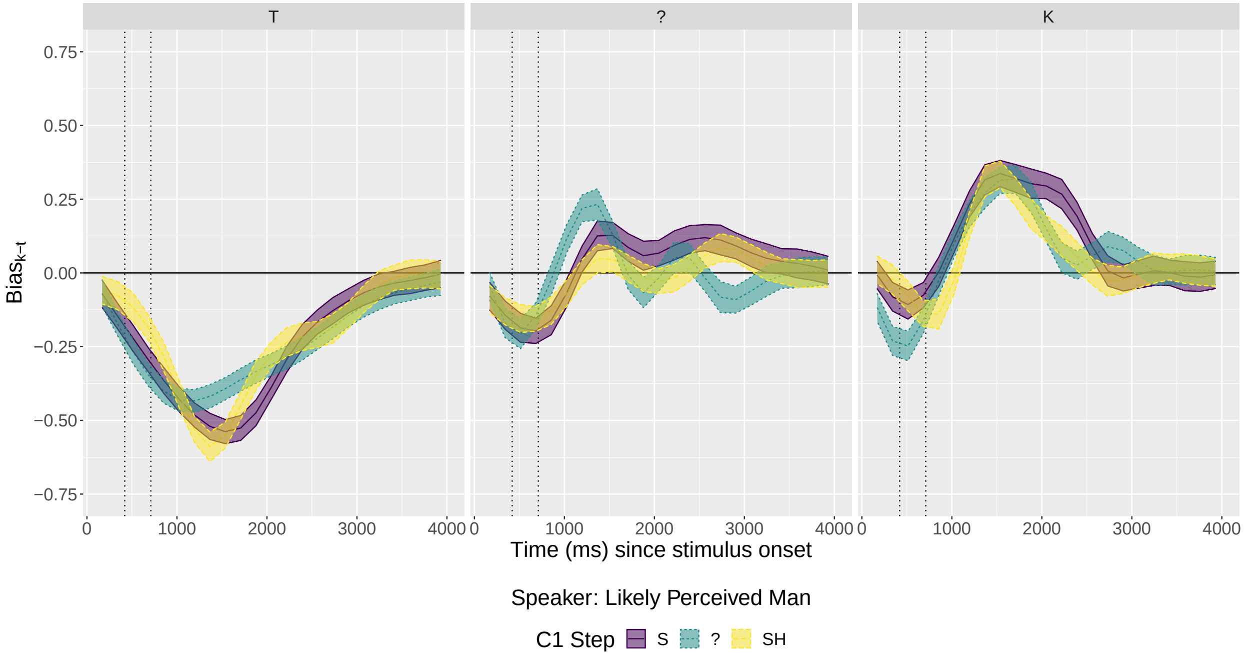


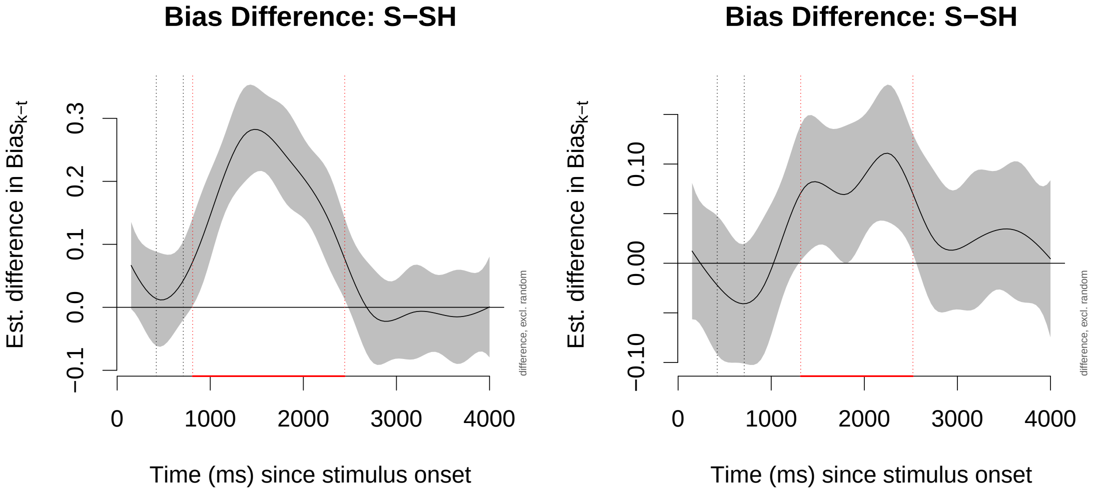
Figure 3.20: Model predicted $Bias_{k-t}$ for the likely perceived man speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

highlighted in red. This significant difference aligns with our Prediction 2 and conforms with the PCCA effect detected in the categorization analyses.

Turning to our more specific Prediction 3, I predicted that this difference would emerge early, being detectable as soon as this information became available, before the onset of C2. This prediction is not supported, as the earliest detectable differences are after the end of the stimulus. The vertical dotted lines in 3.21 correspond to the beginning of C2 and the end of the stimulus, each shifted 150ms later to correspond for the standard accepted delay for programming eye movements. Interestingly, for the likely perceived woman speaker we can observe both an earlier onset (500ms earlier) and greater magnitude (3x peak) of the PCCA effect. I return to consider these differences later in the discussion.

Given that we did not find support for our on-line PCCA Prediction 3, we can immediately discount our more specific Prediction 5 in which we posited that this on-line effect would be modulated by the speaker gender manipulation. Instead, we move towards our less-specific Prediction 4, which held that we would observe speaker gender influencing the PCCA behavior of the intermediate C1 fricative: we expect it to induce PCCA behavior like “S” for the likely perceived man and like “SH” for the likely perceived woman speaker.

Figure 3.22 examines the model predictions for the intermediate C1 - intermediate



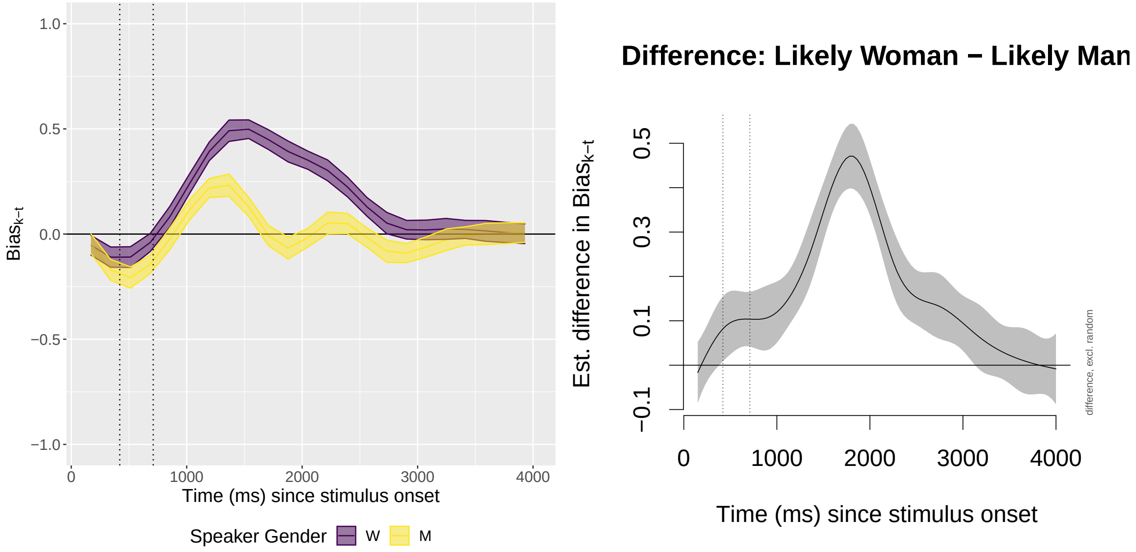
(a) “likely perceived woman”:
sig. different from 0 at 810-2440

(b) “likely perceived man”:
sig. different from 0 at 1320-2520

Figure 3.21: Model predicted $Bias_{k-t}$ difference between S and SH C1 steps for intermediate C2 step. The bolded red regions indicate areas in which the Bias difference is significantly different from zero. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

C2 condition, where we expect the consonant identity to be the most ambiguous and therefore malleable to the influence of earlier gender cues. The left panel presents the trajectories for each speaker, while the right panel presents the difference between these two speakers. If our Prediction 4 held, we would expect the intermediate C1 for the likely woman speaker to be perceived as “SH” and therefore have a lower $Bias_{k-t}$ than the likely man speaker (whose intermediate), ultimately leading to a negative difference trajectory in the right panel. Instead, we observe the opposite pattern, indicating that the likely woman speaker induces greater bias towards “K” fixations for this condition than the likely man speaker. This effect emerges during the stimulus presentation, with the difference between speakers’ bias trajectories becoming significantly different from 0 starting at approximately 380ms after the stimuli began. Allowing for a 150ms delay to program an eye movement, this would correspond with listeners responding to information presented at 130ms: solidly within C1.

As alluded to earlier, it appears that the difference observed here reflects the fact that listeners responding to the likely woman speaker are overall more likely to fixate on (and eventually choose) “K” targets than men for this specific C2 step. Additionally, it is clear that the intermediate C1 step is not as perceptually ambiguous as intended, patterning



(a) Model predicted $Bias_{k-t}$ trajectories by speaker

(b) Trajectory difference measure

Figure 3.22: $Bias_{k-t}$ trajectories for intermediate C1 and C2 steps. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

often with “S”. In order to reliably test Predictions 4-5, we would need to ensure more ambiguous, and perceptually malleable, C1 and C2 tokens.

3.4.2 *alda-arga* class

3.4.2.1 Raw Data

I consider now the fixation data for the *alda-arga* trials. Figure 3.23 presents the percentage of fixations in each 100ms bin as calculated from the raw fixation counts. Each panel represents trials where participants made a particular choice. For example, the top left panel represents fixations in trials where listeners ultimately chose “alda”. Unsurprisingly, we see a peak in “alda” fixations in this panel, as represented by the dark purple line representing these fixations. In all four panels, the target that is eventually chosen receives the most fixations, serving as an indicator that the fixation data are sound and reflect participants’ eventual decision.

While our earlier categorization analyses indicate that listeners’ classification behaviors are linked to our experimental manipulations, the relationship is not perfect. As such, I turn away now from an investigation of fixation trajectories by listeners’ ultimate choices, and instead explore how fixations vary as a function of our experimental manip-

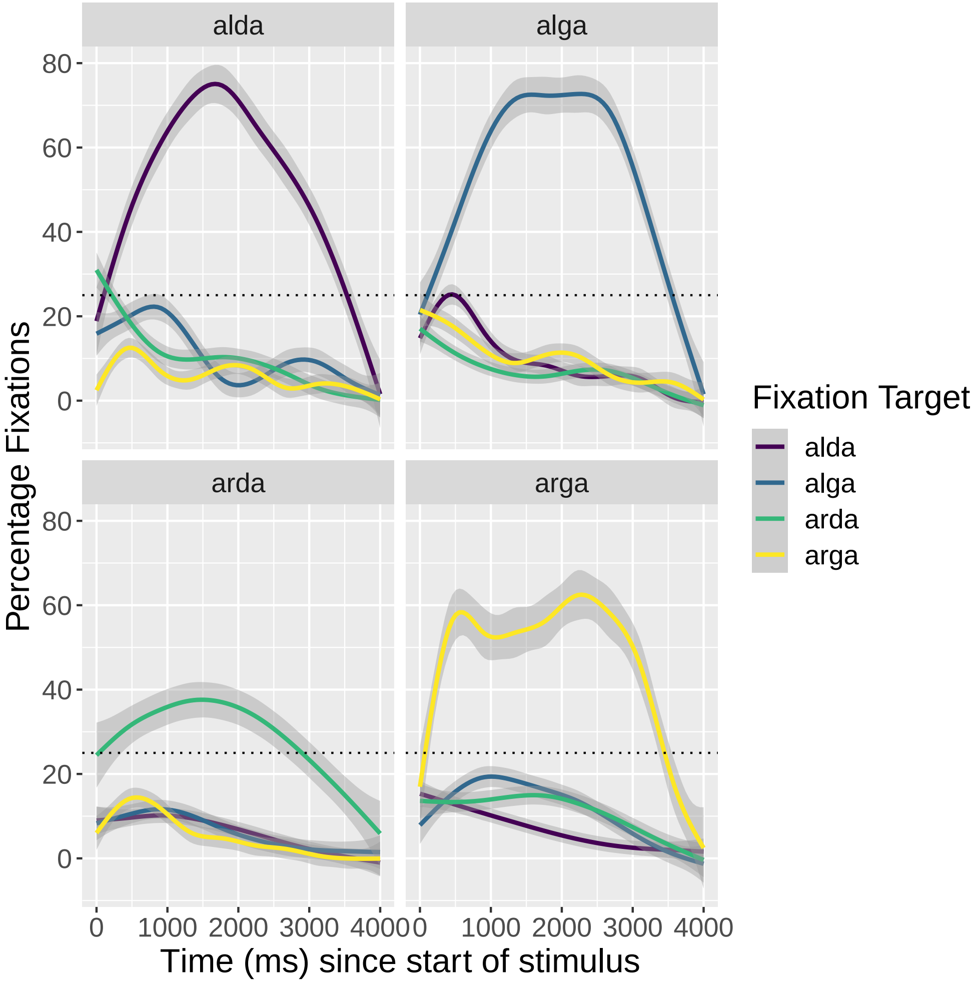


Figure 3.23: Percentage of fixations on specific *alda-arga* Fixation Targets, broken into panels representing listeners' final decisions.

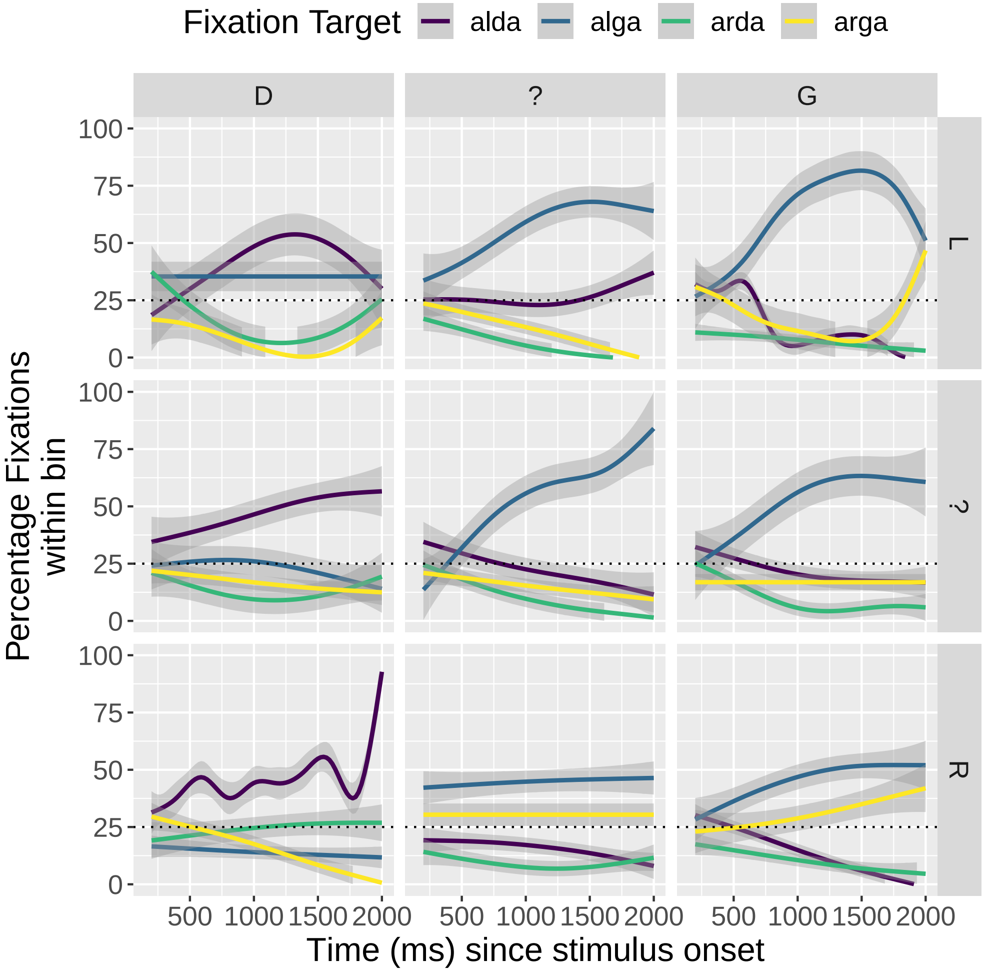


Figure 3.24: Smoothed fixation percentage by C1 and C2.

ulation of C1, C2, and speaker gender.

As alluded to by the classification analysis for this class, participants responses to each C1-C2 condition were less uniform for the *alda-arga* trials than for the *asta-ashka* trials. This is evident in the raw fixation trajectories as well, with many conditions showing longer influence of competitors and less consensus.

Calculated identically to the $Bias_{k-t}$ discussed earlier, the $Bias_{g-d}$ measure represents the relative strength of bias towards fixating on "G" versus on "D". Figure 3.25 presents the trajectory of this bias measure over time for different C2 steps. It is clear that, averaging across other conditions, our C2 manipulations induced changes in fixations in the predicted direction. Unfortunately, the intermediate C2 step appears to be not perceptually intermediate, and instead patterns similarly to "G". This is not unexpected, given similar results for this C2 step in the categorization analyses.

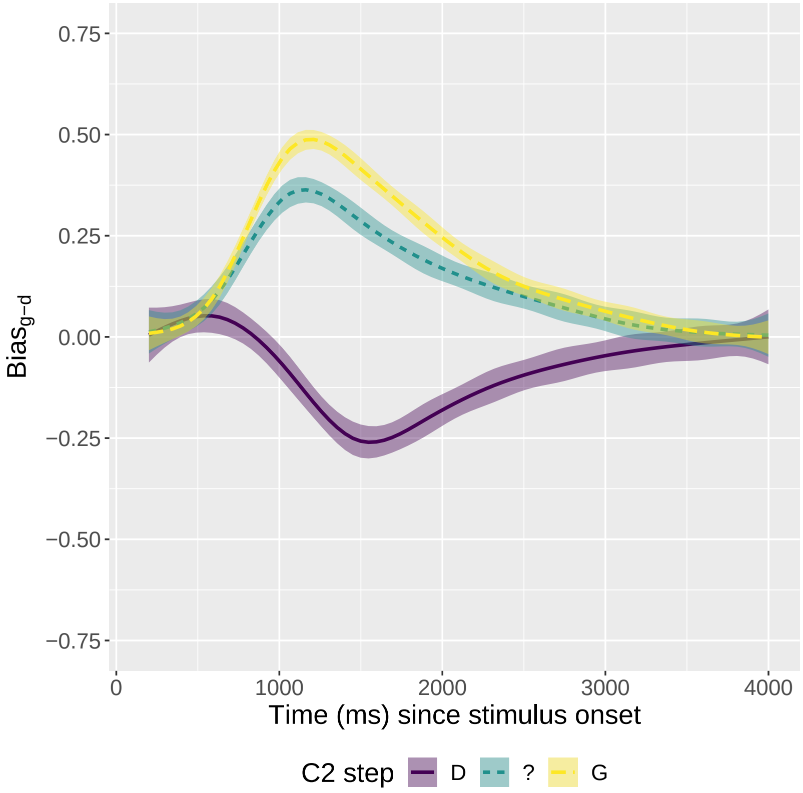


Figure 3.25: $Bias_{g-d}$ averaging across conditions.

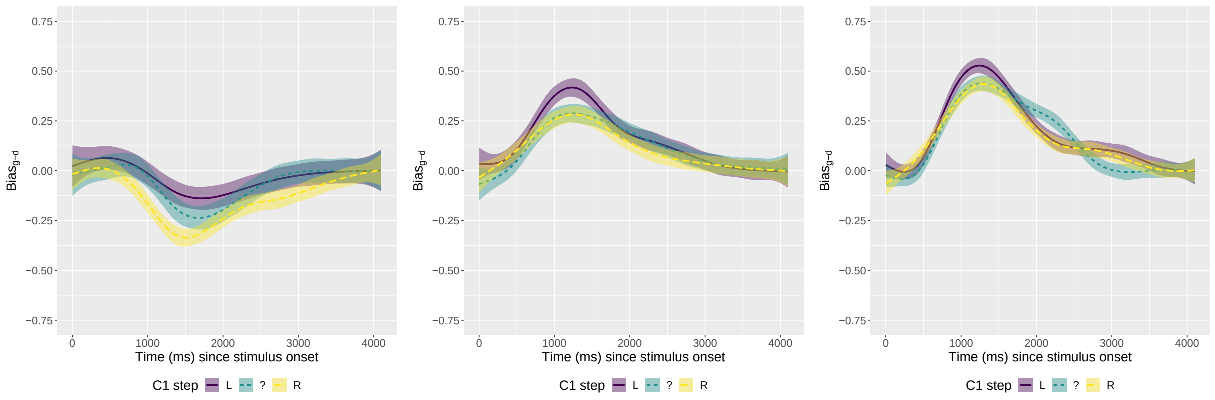
Figure 3.26 explores how this bias evolves over time for each of the various C1 manipulations. In this visualization we observe trends suggestive of our PCCA Prediction 2: a greater bias towards fixating on "G" following "L" than following "R". Like the intermediate step C2, the intermediate step C1 appears to pattern largely with "R", rather than being fully perceptually intermediate.

Next, I test whether these trends persist across different speaker conditions and remain robust when subjected to statistical analysis.

3.4.2.2 Modeling

The modeling process for *alda-arga* trials was identical to that of *asta-ashka* and again our best fitting model includes a smooth of start bin time by the three-way interaction between our fixed effects.

Figure 3.27 presents the model predicted $Bias_{g-d}$ over time for the likely perceived woman speaker. Recall that if two curves (or a curve and a value) do not overlap at a specific time, then the difference between those two curves is significant. For C2 steps "?" and "G", we see early evidence for a bias towards "G", with bias trajectories becoming



(a) C2 - D

(b) C2 - ?

(c) C2 - G

Figure 3.26: $Bias_{g-d}$ separated by C1 and C2 steps.

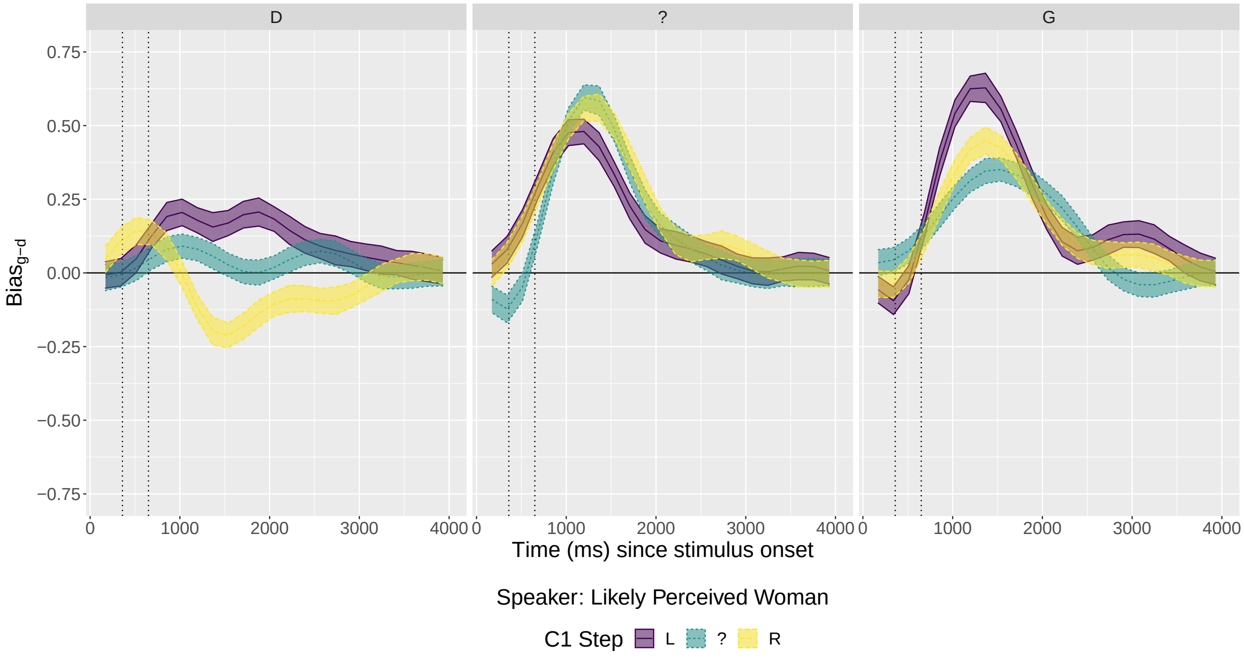


Figure 3.27: Model predicted $Bias_{g-d}$ for the likely perceived woman speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

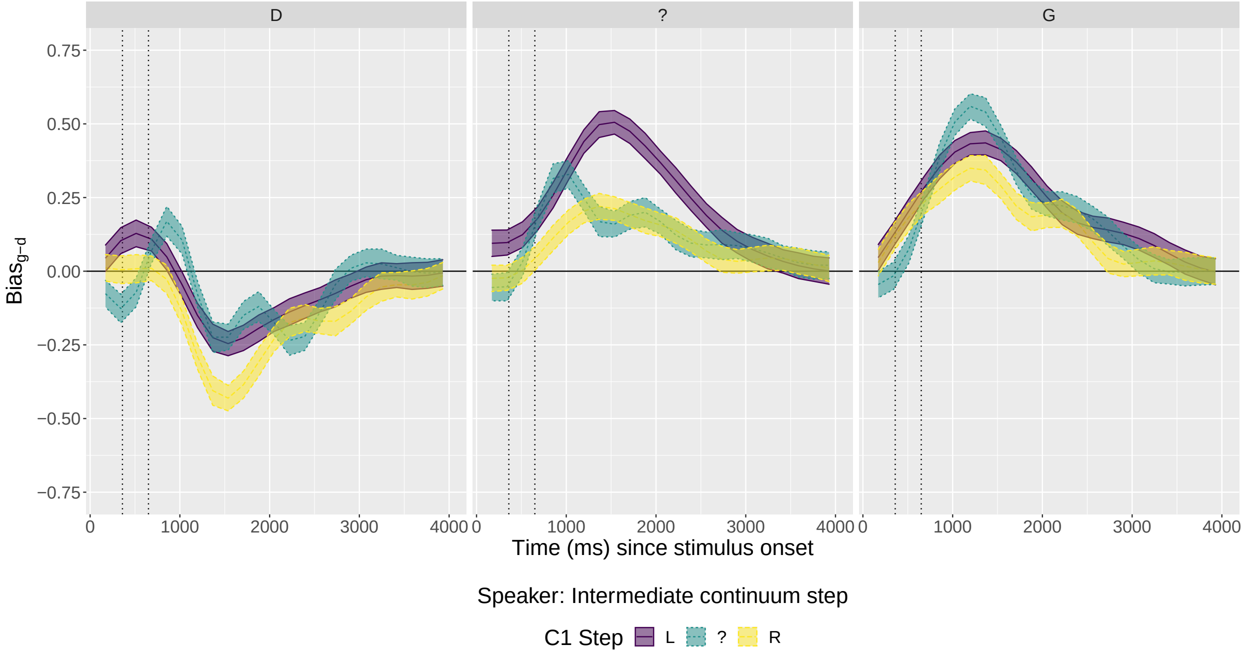


Figure 3.28: Model predicted $Bias_{g-d}$ for the intermediate step speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

significantly greater than 0 during, or slightly after, the stimulus itself. The situation is less clear for the "D" C2 step, where the bias hovers closer to 0 indicating that listeners fixations are more evenly split. This aligns with the categorization analyses, where this "D" token was rated just as often "G" for this speaker. Despite this attenuated bias effect, we can still observe the predicted effects of PCCA: the bias towards "G" is significantly higher following "L" than following "R".

These patterns continue with the likely perceived man and intermediate step speakers in Figure 3.28-3.29. The negative $Bias_{g-d}$ profile for the "D" C2 step becomes clearer for these two speakers though, indicating that these continua endpoints better match listeners' expectations for a /d/-/g/ for these speakers. Additionally, we see clear evidence for the predicted PCCA effect in all conditions, though the timing and magnitude vary.

Although I made no specific predictions for an interaction of speaker gender on C1-C2 PCCA for the *alda-arga* stimuli, we do observe a fixed effect of speaker gender, as in Figure 3.30. We see a significantly increased bias towards fixating on "G" for the likely woman's voice, compared to the likely man's voice, across all C1 exposure conditions. This effect emerges gradually following the end of the stimulus and reflects participants' fixations and subsequent responses on "G" targets for this speaker.

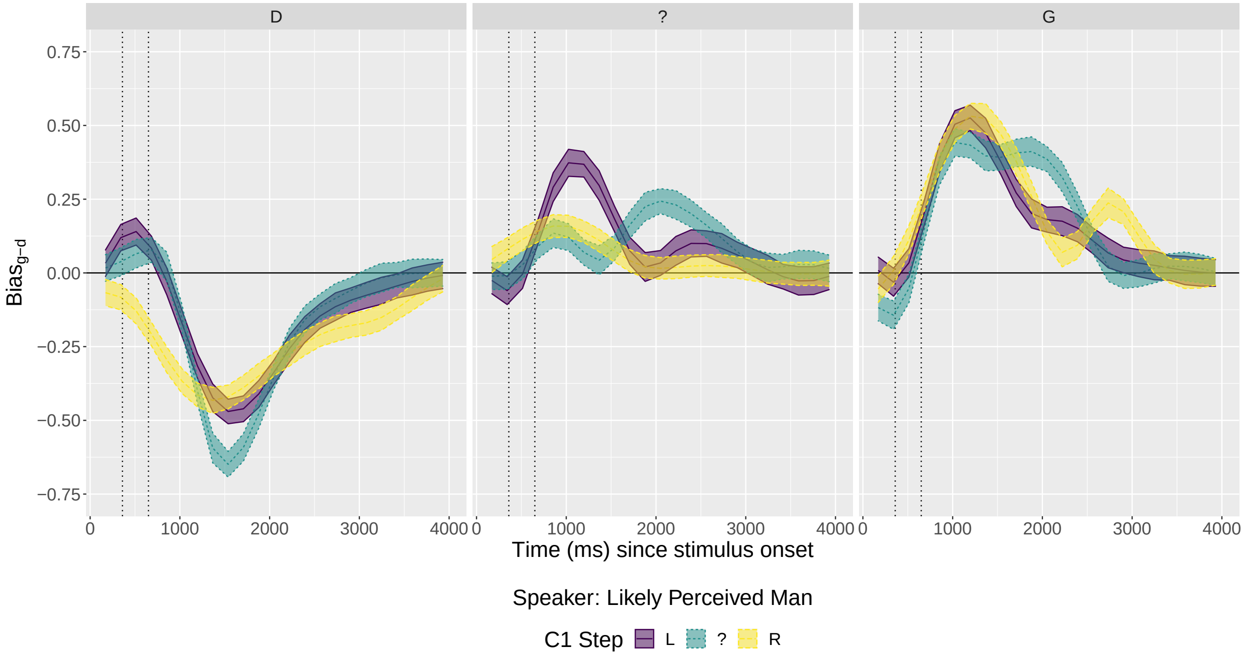


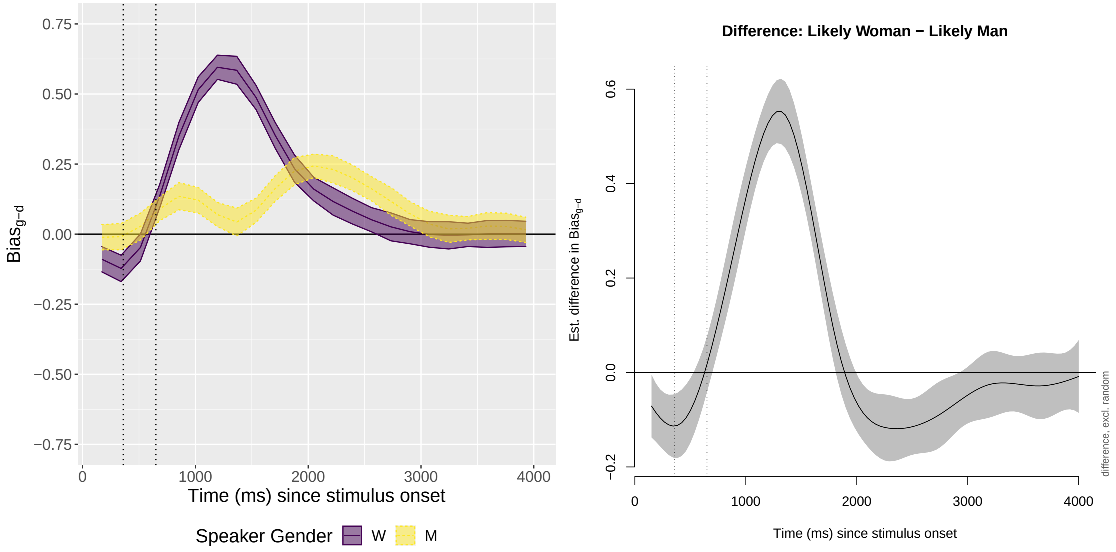
Figure 3.29: Model predicted $Bias_{g-d}$ for the likely perceived man speaker. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

3.4.3 Interim Discussion

As for the *asta-ashka* class, we find clear evidence for the gradual build-up of perceptual compensation for coarticulation in listeners' fixation trajectories for the *alda-arga* stimuli. These patterns align with the categorization data, and demonstrate more fixations to "G" targets when preceded by "L", rather than "R". Additionally, we find evidence an overall effect of speaker (or gender) on C2 fixations, with fixations to "G" being overall more likely for the likely perceived woman speaker. This effect emerges after the end of the stimuli and corresponds to period of peak responses, likely corresponding to a peak in "G" classifications for the likely woman speaker that we observed in the categorization data.

3.5 Chapter Summary and Discussion

This chapter has investigated listeners' perceptual compensation for coarticulation (PCCA) and tested whether this PCCA effect is modulated by the perceived gender of the speaker. Combining listeners' visual fixations during trials with their eventual categorizations, I demonstrated robust PCCA effects for both classes of stimuli: in the *asta-*



(a) Model predicted $Bias_{g-d}$ trajectories by speaker

(b) Trajectory difference measure

Figure 3.30: $Bias_{g-d}$ trajectories for intermediate C1 and C2 steps. The vertical dotted lines indicate the beginning of C2 and the end of the stimulus, shifted 150ms forward to account for delays in programming eye movements.

ashka class, listeners fixated more on “K” and categorized stops more often as “K” when the previous consonant was “S”; in the *alda-arga* class, listeners fixated more on “G” and categorized stops more often as “G” when the previous consonant was “L”.

We did not find support for our original predictions on the role of speaker gender. These predictions centered around the assumption that speaker gender would be sufficient to shift the perception of intermediate C1 steps, and thus change their PCCA influence on the following sounds. We did however observe evidence for fixed speaker-based differences in categorization of and fixations to different C2 targets (e.g., overall more fixations to and categorizations of “K” for women). These fixed effects appear to emerge early in perception, and were detectable in fixations initiated before C2 began.

The lack of support for our original speaker gender predictions could represent the ground truth, and the influence of these social cues does not extend to shifting the PCCA behavior of C1 on C2. This would be support against the view that holds social cues such as this are intrinsically linked to the nature of phonetic representations that are involved in early perceptual processes and acted upon during PCCA effects. However, we cannot reject our original predictions on the basis of these experiments alone, given that we discovered a substantial confound during analysis. Namely, our intermediate consonant steps, despite being intermediate on the synthesized continuum, were not perceptually intermediate. Our key predictions held that for the most ambiguous consonant steps, speaker gender would be able to shift the interpretation and PCCA behavior of consonants in different directions. It appears that all our intermediate steps were far from the perceptual boundary for these continua to be influenced in this way.

This represents a clear line for future work of careful norming of a wider range of steps along these continua in order to choose the true intermediate stimuli steps. The thorough and careful work of Luthra et al. (2021) provides one example of the merit of this approach. With the appropriately ambiguous consonant steps in hand, a more conclusive repetition of the current experiment may be carried out. Should speaker gender be able to induce different directions of PCCA effect of C1 on C2 in this case, it would represent a strong argument for the central role of social cues in this process and the earliest stages of perception.

Chapter 4

Conclusion

Speech perception involves mapping from a highly variable, multi-dimensional signal to an abstract representation of the speaker's intended utterance. Listeners are adept at this task, and can consistently arrive at a correct interpretation of the message despite distractions, variability in the signal, and deteriorated or obscured cues. Previous work has demonstrated that one strategy listeners employ during the speech perception task is taking advantage of the multi-dimensional aspect of the speech signal and drawing upon many informative, and often redundant cues in the signal. One such set of informative cues are the social characteristics of a speaker or community, which can become correlated with particular linguistic variables and which listeners can use to guide their perception. This dissertation investigated the extent to which social cues are recruited during speech perception, focusing on determining whether these cues are mainly active during later decision stages of perception or whether they are also recruited during the earliest stages of perceptual activity.

First, I investigated how social cues interacted with other acoustic cue during a series of selective adaptation experiments. In this experimental paradigm, listeners are exposed to repeated instances of clear exemplars of a category which typically results in a decrease in later categorizations of a stimulus continuum as being members of the exposure category (e.g., being exposed to many clear instances of /ʃ/ leading to less "SH" responses to a /ʃ/-/s continuum). Previous work has demonstrated that lexical frames (e.g., in instances of the Ganong effect) can bias ambiguous stimuli to induce selective adaptation differentially, causing adaptive shifts consistent with the phoneme associated with the lexical frame. This leads to the prediction that social cue information, if it is active at the earliest stages of perception which selective adaptation is argued to occur at, should also be able to guide the interpretation and thus the selective adaptive behavior of ambiguous sounds. While the series of experiments show robust replications of the established selective adaptation effect, we do not observe evidence for the influence of social cues (either acoustic cues to gender or sexuality, or visual face gender cues) on selective adaptation behavior.

Next, I presented categorization and real-time eye-tracking data from a compensation for coarticulation experiment in which classical effects of perceptual compensation for coarticulation (PCCA) were replicated. For example, listeners consistently categorized stimuli more often as /k/ than /t/ when the token was preceded by /s/ rather than /ʃ/. Additionally, these patterns were observed in real-time, as listeners fixated to PCCA-consistent targets quicker and more often than targets not consistent with the PCCA effect. Given the obtained perceptual influence of the first consonant on the second consonant, our critical prediction was that social cues would mediate the perception and subsequent PCCA influences of our first consonant. These perceptual influences would be detected indirectly via the PCCA influence on the second consonant categorizations, thus providing evidence for the early and influential role of this social information. Like the selective adaptation experiments, however, we did not observe evidence for the influence of social cues (here, acoustic gender information) on the PCCA effect.

Taken as a whole, these experimental results provide support for the position that sociophonetic cues may be restricted in their influence to later decision stages of perception, rather than earlier stages of perception. This interpretation rests on the assumption that the critical difference between categorization on the one hand and compensation for coarticulation and selective adaptation effects on the other is the stage at which their effects are active. An alternative account, say in the nature of representations drawn upon in each effect, that separates these sets of experiments would also be consistent with the empirical results presented here. Further experimentation replicating and extending the research carried out here will provide further support and evidence for the role of sociophonetic cues in perception and continue to inform and shape our models of phonetic knowledge and perception.

Bibliography

Abramson, A. S. & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in honor of Peter Ladefoged* (pp. 25–33). New York: Academic Press.

Adams, R. B. & Kveraga, K. (2015). Social Vision: Functional Forecasting and the Integration of Compound Social Cues. *Review of philosophy and psychology*, 6(4), 591–610.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, 183(3), 399–404.

Andreeva, B., Demenko, G., Möbius, B., Zimmerer, F., Jügler, J., & Oleskowicz-Popiel, M. (2014). Differences in pitch profiles in Germanic and Slavic languages. In *INTERSPEECH*.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11), 1174–1220.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. *Frontiers in Human Neuroscience*, 9.

Basu Mallick, D., F. Magnotti, J., & S. Beauchamp, M. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, 22(5), 1299–1307.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, 85(4), 785–821.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, 30(4), 591–627.

Beddor, P. S. & Krakow, R. A. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America*, 106(5), 2868–2887.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, 133(4), 2350–2366.

Beddor, P. S. & Onsuwan, C. (2003). Perception of Prenasalized Stops. *15th International Congress of the Phonetic Sciences*.

Berglund-Barraza, A., Tian, F., Basak, C., & Evans, J. L. (2019). Word Frequency Is Associated With Cognitive Effort During Verbal Working Memory: A Functional Near Infrared Spectroscopy (fNIRS) Study. *Frontiers in Human Neuroscience*, 13.

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatio-temporal Dynamics of Audiovisual Speech Processing. *NeuroImage*, 39(1), 423–435.

Bertelson, P., Vroomen, J., & Gelder, B. d. (2003). Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect. *Psychological Science*.

Bidelman, G. M., Brown, B., Mankel, K., & Nelms Price, C. (2020). Psychobiological Responses Reveal Audiovisual Noise Differentially Challenges Speech Recognition:. *Ear and Hearing*, 41(2), 268–277.

Bonnefond, M., Kastner, S., & Jensen, O. (2017). Communication between Brain Areas Based on Nested Oscillations. *eNeuro*, 4(2).

Bouavichith, D. A. (2019). THE ROLE OF SOCIOINDEXICAL EXPECTATION IN THE PERCEPTION OF GAY MALE SPEECH. *Proceedings of the 19th International Congress of Phonetic Sciences*.

Bouavichith, D. A., Calloway, I., Craft, J. T., Hildebrandt, T., Tobin, S. J., & Beddor, P. S. (2019). PERCEPTUAL INFLUENCES OF SOCIAL AND LINGUISTIC PRIMING ARE BIDIRECTIONAL. *Proceedings of the 19th International Congress of Phonetic Sciences*.

Boyczuk, J. P. & Baum, S. R. (1999). The Influence of Neighborhood Density on Phonetic Categorization in Aphasia. *Brain and Language*, 67(1), 46–70.

Browman, C. P. & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, 49(3-4), 155–180.

Burgering, M. A., Laarhoven, T. v., Baart, M., & Vroomen, J. (2020). Fluidity in the perception of auditory speech: Cross-modal recalibration of voice gender and vowel identity by a talking face:. *Quarterly Journal of Experimental Psychology*.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5), 1–54.

Bürkner, P.-C. & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, 73(3), 420–451.

Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, 11(12), 1110–1123.

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech*, 86(1), 52–68.

Chodroff, E. & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4(s2).

Cibelli, E. S., Leonard, M. K., Johnson, K., & Chang, E. F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and language*, 147, 66–75.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809.

Clopper, C. G. & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32(1), 111–140.

Clopper, C. G. & Pisoni, D. B. (2006). Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language variation and change*, 18(2), 193–221.

Coetzee, A. W., Beddor, P. S., & Wissing, D. P. (2014). Emergent tonogenesis in afrikaans. *The Journal of the Acoustical Society of America*, 135(4), 2421–2422.

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

Dalmaijer, E. & Van der Stigchel, S. (2014). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavioral Research*, 46, 913–921.

Dell, G. S. & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production*. Berlin, New York: DE GRUYTER MOUTON.

Díaz-Campos, M. & Zahler, S. L. (2018). Testing Formal Accounts of Variation: A Sociolinguistic Analysis of Word Order in Negative Word + ms Constructions. *Hispania*, 101(4), 605–619.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, 55(1), 149–179.

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley girls and california vowels. *Journal of Sociolinguistics*, 19(2), 241–256.

D'Onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*, 47(4), 513–539.

Drager, K. (2010). Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass*, 4(7), 473–480.

Drager, K. (2011). Speaker age and vowel perception. *Language and Speech*, 54(1), 99–121.

Drager, K. & Kirtley, M. J. (2016). Awareness, Salience, and Stereotypes in Exemplar-Based Models of Speech Production and Perception. In A. M. Babel (Ed.), *Awareness and Control in Sociolinguistic Research* (pp. 1–24). Cambridge: Cambridge University Press.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, 157(1), 1–42.

Eimas, P. D. & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4, 99–109.

Ellis, L. & Hardcastle, W. J. (2002). Categorical and gradient properties of assimilation in alveolar to velar sequences: evidence from EPG and EMA data. *Journal of Phonetics*, 30(3), 373–396.

Elman, J. L. & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2), 143–165.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, 37(4), 452–465.

Eukel, B. (1980). Phonotactic basis for word frequency effects: Implications for lexical distance metrics. *The Journal of the Acoustical Society of America*, 68(S1), S33–S33.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116, 752–782.

Fiebach, C. J., Friederici, A. D., Mller, K., & von Cramon, D. Y. (2002). fMRI evidence for dual routes to the mental lexicon in visual word recognition. *Journal of Cognitive Neuroscience*, 14(1), 11–23.

Foulkes, P. & Hay, J. B. (2015). The emergence of sociophonetic structure. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 292–313). John Wiley & Sons.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68(2), 161–177.

Fowler, C. A. & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4), 742–754.

Fox, R. A. & Nissen, S. L. (2005). Sex-related acoustic changes in voiceless English fricatives. *Journal of Speech, Language, and Hearing Research*, 48, 753–765.

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, 124(2), 1234–1251.

Fuchs, S. & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent sounds: an interdisciplinary guide* (pp. 281–302). Mouton de Gruyter, 1 edition.

Galle, M. E., Klein-Packard, J., Schreiber, K., & McMurray, B. (2019). What Are You Waiting For? Real-Time Integration of Cues for Fricatives Suggests Encapsulated Auditory Memory. *Cognitive Science*, 43, e12700.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), 110–125.

Garrett, A. & Johnson, K. (2013). Phonetic bias in sound change. In A. C. L. Yu (Ed.), *Origins of Sound Change: Approaches to phonologization*. Oxford University Press.

Gianakas, S. P. & Winn, M. (2016). Exploiting the Ganong effect to probe for phonetic uncertainty resulting from hearing loss. *The Journal of the Acoustical Society of America*, 140(4), 3440–3441.

Gnevsheva, K. (2018). The expectation mismatch effect in accentedness perception of Asian and Caucasian non-native speakers of English. *Linguistics*, 56(3), 581–598.

Gosselin, P. A. & Gagn, J.-P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology*, 50(11), 786–792.

Gow, D. W. & McMurray, B. (2007). Word recognition and phonology. In J. S. Cole & J. Hualdo (Eds.), *Papers in Laboratory Phonology*, volume 9 (pp. 173–200). New York: Mouton de Gruyter.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: A Granger causality analysis of MEG and EEG source estimates. *NeuroImage*, 43(3), 614–623.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28(4), 267–283.

Grosvald, M. & Corina, D. (2012). Perception of long-distance coarticulation: An event-related potential and behavioral study. *Applied Psycholinguistics*, 33(1), 55–82.

Hackett, T. A. (2015). Anatomic organization of the auditory cortex. In G. G. Celesia & G. Hickok (Eds.), *Handbook of Clinical Neurology: The Human Auditory System*, volume 129 of *3rd* (pp. 27–53). Elsevier.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, 28(12), 1860–1871.

Hardcastle, W. J. & Hewlett, N. (1999). *Coarticulation: Theory, Data and Techniques*. Cambridge University Press.

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

Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLOS ONE*, 14(2), e0210793.

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

Helfer, K. S. & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, 117(2), 842–849.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2020). lab.js: A free, open, online study builder.

Hertrich, I., Mathiak, K., Lutzenberger, W., & Ackermann, H. (2009). Time Course of Early Audiovisual Interactions during Speech and Nonspeech Central Auditory Processing: A Magnetoencephalography Study. *Journal of Cognitive Neuroscience*, 21(2), 259–274.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, 108(6), 3013–3022.

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, 8(1), 12372.

Holt, L. L. & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, 119(5), 3059–3071.

House, A. S. & Fairbanks, G. (1953). The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels. *The Journal of the Acoustical Society of America*, 25(1), 105–113.

Howes, D. (1957). On the Relation between the Intelligibility and Frequency of Occurrence of English Words. *Journal of the Acoustical Society of America*, 29(2), 296–305.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5), 1217–1232.

Idemaru, K. & Holt, L. L. (2011). Word Recognition Reflects Dimension-based Statistical Learning. *Journal of Experimental Psychology. Human Perception and Performance*, 37(6), 1939–1956.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, 132(6), 3950–3964.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1–15.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Blackwell Publishers.

Johnson, K. (2020). The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology*, 11(1).

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3), 1252–1263.

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication*, 77, 84–100.

Kang, Y. & Han, S. (2013). Tonogenesis in early Contemporary Seoul Korean: A longitudinal case study. *Lingua*, 134, 62–74.

Kawahara, H. & Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana*, 36(5), 713–727.

Kent, R. D. & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, 74, 74–97.

Kim, J. (2018). *Socially-conditioned links between words and phonetic realizations*. Doctoral Dissertation, University of Hawai'i at Mnoa.

Kim, J. & Drager, K. (2017). Sociophonetic Realizations Guide Subsequent Lexical Access. In *Interspeech 2017* (pp. 621–625).: ISCA.

Kim, J. & Drager, K. (2018). Rapid Influence of Word-Talker Associations on Lexical Access. *Topics in Cognitive Science*, 10(4), 775–786.

Kingston, J. (2011). Tonogenesis. In *The Blackwell Companion to Phonology* (pp. 1–30). American Cancer Society.

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology: Human Perception and Performance*, 42(12), 1969–1988.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Acoustical Society of America*, 67(3), 971–995.

Kleinschmidt, D. F. & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203.

Kong, E. J. & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, 59, 40–57.

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in houston, texas. *University of Pennsylvania Working Papers in Linguistics*, 14(2), 93–101.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Amsterdam: Academic Press is an imprint of Elsevier, 2nd ed. edition.

Kuehn, D. P. & Moll, K. L. (1972). Perceptual Effects of Forward Coarticulation. *Journal of Speech and Hearing Research*, 15(3), 654–664.

Labov, W. (2001). *Principles of Linguistic Change. Volume II: Social Factors*. Oxford: Blackwell.

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89(1), 30–65.

Lam, B. P. W., Xie, Z., Tessmer, R., & Chandrasekaran, B. (2017). The Downside of Greater Lexical Influences: Selectively Poorer Speech Perception in Noise. *Journal of speech, language, and hearing research: JSLHR*, 60(6), 1662–1673.

Lee, H., Politzer-Ahles, S., & Jongman, A. (2013). Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *Journal of Phonetics*, 41(2), 117–132.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3), 1455–1468.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications*, 7(1), 1–9.

Levon, E. (2011). Teasing apart to bring together: gender and sexuality in variationist research. *American Speech*, 86(1), 69–84.

Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.

Lipski, S. C., Escudero, P., & Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, 49(5), 638–650.

Lisker, L. (1986). Voicing in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, 29(1), 3–11.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, 62(3), 615–625.

Luce, P. A. & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1), 1–36.

Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust Lexically Mediated Compensation for Coarticulation: Christmash Time Is Here Again. *Cognitive Science*, 45(4).

Lüttke, C. S., Ekman, M., van Gerven, M. A. J., & de Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports*, 6(1), 1–7.

Mack, S. & Munson, B. (2012). The influence of /s/ quality on ratings of men's sexual orientation: explicit and implicit measures of the 'gay lisp' stereotype. *Journal of Phonetics*, 40, 198–212.

Magnotti, J. F. & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLOS Computational Biology*, 13(2), e1005229.

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*, 4.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: the ghost of Christmash past. *Cognitive Science*, 27(2), 285–298.

Mann, V. & Repp, B. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, 28(3), 213–218.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, 28(5), 407–412.

Martin, J. G. & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *The Journal of the Acoustical Society of America*, 69(2), 559–567.

Massaro, D. W. & Cohen, M. M. (1977). Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics*, 22(4), 373–382.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44, 314–324.

Mattys, S. L. & Scharenborg, O. (2014). Phoneme categorization and discrimination in younger and older adults: A comparative analysis of perceptual, lexical, and attentional factors. *Psychology and Aging*, 29(1), 150–162.

Mattys, S. L. & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, 65(2), 145–160.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86.

McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, 62(1), 115–136.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language & Speech*, 58(4), 502–521.

McGowan, K. B. & Babel, A. M. (2019). Perceiving isn't believing: Divergence in levels of sociolinguistic awareness. *Language in Society*, (pp. 1–26).

McGurk, H. & MacDonald (1976). Hearing lips and seeing voices | Nature. *Nature*, 264, 746–748.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, 15(6), 1064–1071.

McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2), 219–246.

Ménard, L., Toupin, C., Baum, S. R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America*, 134(4), 2975–87.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (New York, N.Y.)*, 343(6174), 1006–1010.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics*, 68(7), 1227–1240.

Mitterer, H. & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, 69(4), 527–545.

Munson, B. (2011). The influence of actual and imputed talker gender on fricative perception, revisited. *The Journal of the Acoustical Society of America*, 130(5), 2631–2634.

Munson, B. & Babel, M. (2007). Loose lips and silver tongues, or, projecting sexual orientation through speech. *Language and Linguistics Compass*, 1(5), 416–449.

Munson, B., Jefferson, S. V., & McDonald, E. C. (2006a). The influence of perceived sexual orientation on fricative identification. *The Journal of the Acoustical Society of America*, 119(4), 2427–2437.

Munson, B., McDonalad, E. C., DeBoe, N. L., & White, A. R. (2006b). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, 34, 202–240.

Munson, B., Ryherd, K., & Kemper, S. (2017). Implicit and explicit gender priming in english lingual sibilant fricative perception. *Linguistics*, 55(5), 1073–1107.

Myers, E. B. & Blumstein, S. E. (2008). The neural bases of the lexical effect: an fMRI investigation. *Cerebral Cortex (New York, N.Y.: 1991)*, 18(2), 278–288.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3), 873–889.

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

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

Nittrouer, S. & Lowenstein, J. H. (2009). Does harmonicity explain childrens cue weighting of fricative-vowel syllables? *The Journal of the Acoustical Society of America*, 125(3), 1679–1692.

Noe, C. & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition*, 197, 104162.

Norris, D. & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3), 299–325.

Ohala, J. J. (1989). Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (Eds.), *Language change: Contributions to the study of its causes* (pp. 173–198). Mouton de Gruyter.

Olasagasti, I., Bouton, S., & Giraud, A.-L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex*, 68, 61–75.

Peltola, M. S., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., & Ntnen, R. (2003). Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience Letters*, 352(1), 25–28.

Pickles, J. O. (2012). *Introduction to the Physiology of Hearing*. ProQuest Ebook Central.

Pickles, J. O. (2015). Auditory pathways: anatomy and physiology. In G. G. Celesia & G. Hickok (Eds.), *Handbook of Clinical Neurology: The Human Auditory System*, volume 129 of *3rd* (pp. 3–25). Elsevier B. V.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech communication*, 4(1-3), 75–95.

Pitt, M. A. & Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1120–1135.

Podesva, R. J. (2006). *Phonetic detail in sociolinguistic variation: its linguistic significance and role in the construction of social meaning*. PhD thesis, Stanford University.

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

Polka, L. & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *The Journal of the Acoustical Society of America*, 78(4), 1187–1197.

Prabhakaran, R., Blumstein, S. E., Myers, E. B., Hutchison, E., & Britton, B. (2006). An event-related fMRI investigation of phonologicallexical competition. *Neuropsychologia*, 44(12), 2209–2221.

Pratt, H., Bleich, N., & Mittelman, N. (2015). Spatiotemporal distribution of brain activity associated with audiovisually congruent and incongruent speech and the McGurk Effect. *Brain and Behavior*, 5(11).

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification:. *Journal of Language and Social Psychology*, 18(1), 10–30.

Raphael, L. J. (2005). Acoustic Cues to the Perception of Segmental Phonemes. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 182–206). Oxford, UK: Blackwell Publishing Ltd.

Reinisch, E. & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, 41(2), 101–116.

Repp, B. H. (1981). Phonetic trading relations and context effects: New evidence for a phonetic mode of perception. *Psychological Bulletin*, 92, 81–110.

Roberts, M. & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, 30(4), 309–314.

Robinson, C. W. & Sloutsky, V. M. (2004). Auditory Dominance and Its Change in the Course of Development. *Child Development*, 75(5), 1387–1401.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1019–1034.

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon:. *Current Directions in Psychological Science*.

Rysling, A., Kingston, J., Staub, A., Cohen, A., & Starns, J. (2015). Early Ganong Effects. *Proceedings of the 18th International Congress of Phonetic Sciences.*

Saalasti, S., Ktsyri, J., Tiippana, K., Laine-Hernandez, M., von Wendt, L., & Sams, M. (2012). Audiovisual Speech Perception and Eye Gaze Behavior of Adults with Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 42(8), 1606–1615.

Sachs, J. (1975). Cues to the identification of sex in children's speech. In B. Thorne & N. Henley (Eds.), *Language & sex: Difference & dominance*. Massachusetts, NH: Newbury House.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, 71(1), 145–163.

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1124–1131.

Samuel, A. G. (1997). Lexical Activation Produces Potent Phonemic Percepts. *Cognitive Psychology*, 32(2), 97–127.

Samuel, A. G. (2001). Knowing a Word Affects the Fundamental Perception of The Sounds Within it:. *Psychological Science*, 12(4).

Samuel, A. G. (2011). Speech Perception. *Annual Review of Psychology*, 62(1), 49–72.

Samuel, A. G. & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychonomic bulletin & review*, 22(6), 1746–1752.

Samuel, A. G. & Lieblich, J. (2014). Visual Speech Acts Differently Than Lexical Context in Supporting Speech Perception. *Journal of experimental psychology. Human perception and performance*, 40(4), 1479–1490.

Schwartz, R. G., Scheffler, F. L. V., & Lopez, K. (2013). Speech perception and lexical effects in specific language impairment. *Clinical Linguistics & Phonetics*, 27(5), 339–354.

Shahin, A. J., Bishop, C. W., & Miller, L. M. (2009). Neural mechanisms for illusory filling-in of degraded speech. *NeuroImage*, 44(3), 1133–1143.

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America*, 132(2), EL95–EL101.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, 10(1), 1–9.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cerebral cortex (New York, N.Y. : 1991)*, 17(10), 2387–2399.

Slawinski, E. B. & Fitzgerald, L. K. (1998). Perceptual development of the categorization of the /r-w/ contrast in normal children. *Journal of Phonetics*, 26(1), 27–43.

Sloutsky, V. M. & Napolitano, A. C. (2003). Is a Picture Worth a Thousand Words? Preference for Auditory Modality in Young Children. *Child Development*, 74(3), 822–833.

Sonderegger, M. & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. *Proceedings of the Annual Meeting of the Cognitive Science Society: UC Merced*, 32, 7.

Strand, E. A. & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: results of the 3rd KONVENS conference* (pp. 14–26). Berlin: Mouton de Gruyter.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole & J. I. Hualde (Eds.), *Laboratory Phonology 9* (pp. 65–86). New York, USA: Mouton de Gruyter.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4(1015), 1–13.

Szakay, A., Babel, M., & King, J. (2016). Social categories are shared across bilinguals' lexicons. *Journal of Phonetics*, 59, 92–109.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye Movements and Lexical Access in Spoken-Language Comprehension: Evaluating a Linking Hypothesis between Fixations and Linguistic Processing. *Journal of Psycholinguistic Research*, 29(6), 557–580.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, 125(6), 3974–3982.

Thomas, E. R. & Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics*, 8(1), 54–87.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, 5.

Tobin, S. J., Cho, P. W., Jennett, P. M., & Magnuson, J. S. (2010). Effects of Anticipatory Coarticulation on Lexical Access. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32, 2200–2205.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185, 1–20.

Toscano, J. C. & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, 34(3), 434–464.

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Thoret, H. (2007). Speech and Non-Speech Audio-Visual Illusions: A Developmental Study. *PLOS ONE*, 2(8), e742.

van Linden, S., Stekelenburg, J. J., Tuomainen, J., & Vroomen, J. (2007). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience Letters*, 420(1), 49–52.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). itsadug: Interpreting time series and autocorrelated data using gamms. R package version 2.4.1.

Vitevitch, M. S. & Luce, P. A. (1998). When Words Compete: Levels of Processing in Perception of Spoken Words. *Psychological Science*, 9(4), 325–329.

von der Malsburg, T. (2019). *Saccades: Saccade and Fixation Detection in R*. version 0.2.1. https://github.com/tmalsburg/saccades.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditoryvisual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(3), 572–577.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, 44, 55–61.

Walker, A., García, C., Cortés, Y., & Campbell-Kibler, K. (2014). Comparing social meanings across listener and speaker groups: The indexical field of Spanish /s/. *Language Variation and Change*, 26(2), 169–189.

Walker, A. & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, 2, 219–237.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, 167(3917), 392–393.

Warren, R. M. & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. *Perception & Psychophysics*, 16(1), 150–156.

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, 93(4), 2152–2159.

Whalen, D. H. & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, 237(4811), 169–171.

Wieser, M. J. & Brosch, T. (2012). Faces in Context: A Review and Systematization of Contextual Influences on Affective Face Processing. *Frontiers in Psychology*, 3.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, 131(2), 1465–1479.

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, 4.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3–36.

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative Audiovisual Speech Perception in Noise: Dissociable Effects of the Speaker and Listener. *PLoS ONE*, 9(12).

Yazawa, K., Whang, J., Kondo, M., & Escudero, P. (2019). Language-dependent cue weighting: An investigation of perception modes in L2 learning:. *Second Language Research*.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron*, 102(6), 1096–1110.

Ylinen, S., Uther, M., Latvala, A., Vepslinen, S., Iverson, P., Akahane-Yamada, R., & Ntnen, R. (2009). Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, 22(6), 1319–1332.

Yu, A. C. & Zellou, G. (2019). Individual Differences in Language Processing: Phonology. *Annual Review of Linguistics*, 5(1), 131–150.

Zatorre, R. J. & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, 11(10), 946–953.

Zellou, G. & Dahan, D. (2019). Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in English. *Journal of Phonetics*, 76, 100910.

Zheng, Y. & Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attention, Perception, & Psychophysics*, 79(6), 1841–1859.

Appendix A

Experiment	**Materials/Preregistration**
Selective Adaptation Exp. 1	`https://osf.io/2zj97/`
Selective Adaptation Exp. 2	`https://osf.io/r5wta/`
Selective Adaptation Exp. 3	`https://osf.io/m4dy7/`
Selective Adaptation Exp. 4	`https://osf.io/brx5h/`
Selective Adaptation Exp. 5	`https://osf.io/xdtwr/`
Eyetracking Experiment	`https://osf.io/edgxq/`

Table A.1: Experiment Materials and Preregistrations

Appendix B

	Exposure Gender	Exposure Fricative	.epred	.lower	.upper	.width
1	Intermediate Step	Intermediate Step	0.58	0.52	0.65	0.66
2	Intermediate Step	Likely Perceived ″S″	0.63	0.56	0.69	0.66
3	Intermediate Step	Likely Perceived ″SH″	0.54	0.47	0.61	0.66
4	Likely Perceived Woman	Intermediate Step	0.49	0.41	0.55	0.66
5	Likely Perceived Woman	Likely Perceived ″S″	0.56	0.49	0.62	0.66
6	Likely Perceived Woman	Likely Perceived ″SH″	0.54	0.47	0.62	0.66
7	Likely Perceived Man	Intermediate Step	0.57	0.51	0.64	0.66
8	Likely Perceived Man	Likely Perceived ″S″	0.70	0.64	0.77	0.66
9	Likely Perceived Man	Likely Perceived ″SH″	0.52	0.45	0.59	0.66
10	Intermediate Step	Intermediate Step	0.58	0.47	0.69	0.89
11	Intermediate Step	Likely Perceived ″S″	0.63	0.52	0.75	0.89
12	Intermediate Step	Likely Perceived ″SH″	0.54	0.43	0.65	0.89
13	Likely Perceived Woman	Intermediate Step	0.49	0.37	0.60	0.89
14	Likely Perceived Woman	Likely Perceived ″S″	0.56	0.44	0.66	0.89
15	Likely Perceived Woman	Likely Perceived ″SH″	0.54	0.42	0.66	0.89
16	Likely Perceived Man	Intermediate Step	0.57	0.46	0.67	0.89
17	Likely Perceived Man	Likely Perceived ″S″	0.70	0.59	0.80	0.89
18	Likely Perceived Man	Likely Perceived ″SH″	0.52	0.40	0.63	0.89

Table B.1: Experiment 1 Posterior HDI Estimates for Post-test block (block 6) and Middle Sibilant Step (3)

	Exposure Gender	Exposure Fricative	.epred	.lower	.upper	.width
1	Intermediate Step	Intermediate Step	0.47	0.38	0.55	0.66
2	Intermediate Step	Likely Perceived ”S”	0.66	0.60	0.72	0.66
3	Intermediate Step	Likely Perceived ”SH”	0.53	0.45	0.61	0.66
4	Likely Perceived Woman	Intermediate Step	0.50	0.43	0.57	0.66
5	Likely Perceived Woman	Likely Perceived ”S”	0.65	0.58	0.71	0.66
6	Likely Perceived Woman	Likely Perceived ”SH”	0.39	0.32	0.45	0.66
7	Likely Perceived Man	Intermediate Step	0.71	0.64	0.78	0.66
8	Likely Perceived Man	Likely Perceived ”S”	0.85	0.82	0.90	0.66
9	Likely Perceived Man	Likely Perceived ”SH”	0.52	0.45	0.60	0.66
10	Intermediate Step	Intermediate Step	0.47	0.35	0.63	0.89
11	Intermediate Step	Likely Perceived ”S”	0.66	0.55	0.76	0.89
12	Intermediate Step	Likely Perceived ”SH”	0.53	0.39	0.65	0.89
13	Likely Perceived Woman	Intermediate Step	0.50	0.39	0.61	0.89
14	Likely Perceived Woman	Likely Perceived ”S”	0.65	0.54	0.76	0.89
15	Likely Perceived Woman	Likely Perceived ”SH”	0.39	0.27	0.50	0.89
16	Likely Perceived Man	Intermediate Step	0.71	0.59	0.81	0.89
17	Likely Perceived Man	Likely Perceived ”S”	0.85	0.79	0.91	0.89
18	Likely Perceived Man	Likely Perceived ”SH”	0.52	0.40	0.64	0.89

Table B.2: Experiment 2 Posterior HDI Estimates for Post-test block (block 4) and Middle Sibilant Step (3)

	Exposure Gender	Exposure Fricative	.epred	.lower	.upper	.width
1	Intermediate Step	Intermediate Step	0.49	0.43	0.55	0.66
2	Intermediate Step	Likely Perceived ”S”	0.63	0.57	0.69	0.66
3	Intermediate Step	Likely Perceived ”SH”	0.41	0.35	0.48	0.66
4	Likely Perceived Woman	Intermediate Step	0.54	0.49	0.61	0.66
5	Likely Perceived Woman	Likely Perceived ”S”	0.64	0.57	0.70	0.66
6	Likely Perceived Woman	Likely Perceived ”SH”	0.52	0.45	0.58	0.66
7	Likely Perceived Man	Intermediate Step	0.54	0.48	0.61	0.66
8	Likely Perceived Man	Likely Perceived ”S”	0.70	0.64	0.76	0.66
9	Likely Perceived Man	Likely Perceived ”SH”	0.50	0.44	0.58	0.66
10	Intermediate Step	Intermediate Step	0.49	0.39	0.59	0.89
11	Intermediate Step	Likely Perceived ”S”	0.63	0.53	0.73	0.89
12	Intermediate Step	Likely Perceived ”SH”	0.41	0.31	0.52	0.89
13	Likely Perceived Woman	Intermediate Step	0.54	0.43	0.65	0.89
14	Likely Perceived Woman	Likely Perceived ”S”	0.64	0.52	0.73	0.89
15	Likely Perceived Woman	Likely Perceived ”SH”	0.52	0.40	0.63	0.89
16	Likely Perceived Man	Intermediate Step	0.54	0.42	0.65	0.89
17	Likely Perceived Man	Likely Perceived ”S”	0.70	0.59	0.79	0.89
18	Likely Perceived Man	Likely Perceived ”SH”	0.50	0.40	0.63	0.89

Table B.3: Experiment 3 Posterior HDI Estimates for Post-test block (block 4) and Middle Sibilant Step (3)

	Exposure Gender	Exposure Fricative	.epred	.lower	.upper	.width
1	Intermediate Step	Intermediate Step	0.67	0.61	0.75	0.66
2	Intermediate Step	Likely Perceived ”S”	0.70	0.65	0.78	0.66
3	Intermediate Step	Likely Perceived ”SH”	0.46	0.38	0.52	0.66
4	Likely Perceived Woman	Intermediate Step	0.59	0.52	0.66	0.66
5	Likely Perceived Woman	Likely Perceived ”S”	0.66	0.59	0.74	0.66
6	Likely Perceived Woman	Likely Perceived ”SH”	0.48	0.40	0.55	0.66
7	Likely Perceived Man	Intermediate Step	0.61	0.55	0.69	0.66
8	Likely Perceived Man	Likely Perceived ”S”	0.76	0.71	0.82	0.66
9	Likely Perceived Man	Likely Perceived ”SH”	0.50	0.42	0.58	0.66
10	Intermediate Step	Intermediate Step	0.67	0.55	0.78	0.89
11	Intermediate Step	Likely Perceived ”S”	0.70	0.59	0.81	0.89
12	Intermediate Step	Likely Perceived ”SH”	0.46	0.34	0.58	0.89
13	Likely Perceived Woman	Intermediate Step	0.59	0.47	0.71	0.89
14	Likely Perceived Woman	Likely Perceived ”S”	0.66	0.55	0.78	0.89
15	Likely Perceived Woman	Likely Perceived ”SH”	0.48	0.35	0.61	0.89
16	Likely Perceived Man	Intermediate Step	0.61	0.50	0.73	0.89
17	Likely Perceived Man	Likely Perceived ”S”	0.76	0.66	0.85	0.89
18	Likely Perceived Man	Likely Perceived ”SH”	0.50	0.37	0.63	0.89

Table B.4: Experiment 4 Posterior HDI Estimates for Post-test block (block 4) and Middle Sibilant Step (3)

	Exposure Sexuality	Exposure Fricative	.epred	.lower	.upper	.width
1	Intermediate Step	Intermediate Step	0.47	0.41	0.54	0.66
2	Intermediate Step	Likely Perceived ”S”	0.66	0.60	0.72	0.66
3	Intermediate Step	Likely Perceived ”SH”	0.57	0.50	0.64	0.66
4	Likely Perceived Gay Man	Intermediate Step	0.47	0.40	0.53	0.66
5	Likely Perceived Gay Man	Likely Perceived ”S”	0.66	0.60	0.73	0.66
6	Likely Perceived Gay Man	Likely Perceived ”SH”	0.32	0.26	0.38	0.66
7	Likely Perceived Straight Man	Intermediate Step	0.54	0.48	0.61	0.66
8	Likely Perceived Straight Man	Likely Perceived ”S”	0.73	0.68	0.79	0.66
9	Likely Perceived Straight Man	Likely Perceived ”SH”	0.46	0.39	0.53	0.66
10	Intermediate Step	Intermediate Step	0.47	0.36	0.58	0.89
11	Intermediate Step	Likely Perceived ”S”	0.66	0.55	0.75	0.89
12	Intermediate Step	Likely Perceived ”SH”	0.57	0.46	0.68	0.89
13	Likely Perceived Gay Man	Intermediate Step	0.47	0.36	0.58	0.89
14	Likely Perceived Gay Man	Likely Perceived ”S”	0.66	0.54	0.76	0.89
15	Likely Perceived Gay Man	Likely Perceived ”SH”	0.32	0.22	0.42	0.89
16	Likely Perceived Straight Man	Intermediate Step	0.54	0.43	0.64	0.89
17	Likely Perceived Straight Man	Likely Perceived ”S”	0.73	0.63	0.82	0.89
18	Likely Perceived Straight Man	Likely Perceived ”SH”	0.46	0.35	0.57	0.89

Table B.5: Experiment 5 Posterior HDI Estimates for Post-test block (block 4) and Middle Sibilant Step (3)