The Effect of Listener Experience and Social Expectation on Illusory Percepts

by

Justin T. Craft

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in the University of Michigan
2024

Doctoral Committee:

Professor Patrice Speeter Beddor, Co-Chair
Associate Professor Jonathan R. Brennan, Co-Chair
Associate Professor David Brang
Associate Professor Chandan R. Narayan, York University

Justin T. Craft

juscraft@umich.edu

ORCID iD: 0000-0003-0588-3714

Dedication

This dissertation is dedicated to my great-grandmother June McCoy, grandfather Raymond Hugli, mentor Samuel D. Epstein, and student Peter Hart, in each of whom I saw a small reflection of myself but who all died while I completed this work. Thank you for the time we were able to share with one another, sorry you missed the party.

Acknowledgements

In late December 1997 I was invited to Knott's Berry Farm, a Southern California amusement park, with three or four friends to celebrate the tenth birthday of our friend Jay-Dee Morgan. There wasn't a soul at the park, and, left to our own devices, our group of nine-year-old boys ran at top speed from ride to ride dipping and ducking around stanchions and soaking up the fact that we essentially had the park all to ourselves. That's when we got to the front of one line only to be met by the sight of 70 or 80 towering grown men. Through short, often one-sided, conversations in line and the experience of lap bars that never came *all* the way down when sharing a roller coaster car with a man five times our size, we learned that they were the University of Michigan football team (the team that would soon be national champions) enjoying some time in Southern California before playing in the 1998 Rose Bowl. Like young boys are wont to do, we followed them around from ride to ride and were likely a nuisance they hoped to lose throughout the day. At some point during that day for the first time in my life I thought to myself that I wanted to go to college. I came home and told my parents how great the guys from Michigan were and how badly I too wanted to be a Wolverine. My dad bought matching blue hats with a maize block 'M' on the front for the two of us, and I sat mesmerized in front of the TV on New Year's Day 1998 as their winged helmets raced across the screen. They beat Washington State 21 – 16. All this to say, "Thanks, Jay Dee." Maybe ending up here in Ann Arbor was fate all along but I like to think you helped set something in motion by inviting me along to celebrate your birthday 26 years ago.

The research interests that have blossomed at Michigan were planted years before and were fostered by a host of caring friends and mentors without whom this dissertation would likely never have happened. Thank you to Natalie Schiavone for handing me the first text I ever read that mentioned the term linguistics and challenging me to sit with the concepts and apply them to my writing at Long Beach City College. Likewise, thank you to Ken Borgers who taught me everything I know about radio production and audio recording. When you take away the transmitter atop the radio station, it turns out that it's just a phonetics laboratory in the end. Thank you to Bruce Hayes for setting the fire in me when I first got to UCLA and thank you to Kie Zuraw, Megha Sundara, Pat Keating, Laura McPherson, Laura Kalin, Adam Chong, and Dustin Bowers for keeping it burning via your instruction, guidance, and friendship. Thank you to the members of my 'first cohort' at UCLA, in particular Adam King, John DeMartini, Ann Aly, and Sözen Özkan for your camaraderie and continued friendship. To Dustin Feldt, thank you for the shared bi-weekly Metro rides during my M.A. to the Staples Center to see the Kings where I'd sketch out OT tableaux or syntax trees on the backs of receipts and napkins when you'd ask what I was working on or learning. Those conversations were absolutely foundational for how I've approached teaching new concepts to students and communicating my work to colleagues, peers, reviewers, and anyone who will listen. Thank you to Tony Natoci for suggesting I put in, "one more application and why not Michigan!?" and to Jennifer Hoffseth, Anna Garbier, Rachel Jacobson, Erin Boon, Gunnar Lund, Olivia Redfield, Kevin Steen, Papa Ngom, Tuomas Paajanen, Daan van Esch, Jeremy O'Brien and the rest of the SDO and localization teams at Google for showing me where my knowledge about linguistics could be applicable outside of the academy – it's made walking away from the academic market far less terrifying. To Pud, you almost got me to Alberta, bud. I just couldn't do the 3:45 sunsets.

I could not have been luckier to research and study at the University of Michigan and I've had the privilege to grow alongside and learn from some of the most brilliant graduate students I've ever met. Each of them has had a lasting impact on how I view my work and where it sits within the field of linguistics. Thank you first and foremost to the P-Boiz: Dominique Bouavichith, Ian Calloway, and Ruaridh Purse. You three have been the definition of iron sharpens iron. At the end of the day, I was never working harder or thinking more deeply than when I was working alongside each of you. I'm worse off in whatever I do next without you pushing me. To Carrie Ann Morgan, thanks for handholding a stubborn generativist like me into the world of linguistic anthropology, I couldn't have done a lot of what I have without your camaraderie, care, and guidance. To Kelly Wright, thank you for always pushing me one step further and reminding me that my work mattered when I felt like it didn't. To the members of Pam's lab group thank you for the bi-weekly feedback and for always challenging me to think of problems beyond my own. To Dani Burgess, Dominique Canning, Demet Kayabaşı, and Kate Sherwood thank you for all the board games nights and trivia wins at Bløm. You all kept me sane throughout the journey and I couldn't be more thankful for your friendship.

I also wouldn't be where I am today without the instruction and guidance of some key faculty members here at Michigan. To Sam Epstein, I miss you every day, buddy. When Augs gets in trouble I use his middle name "August Samuel Craft" but when he's in *real* trouble it's "August Samuel David Epstein Craft". Can't wait to tell him about you and I hope he turns out just as curious about the world. Robin, thank for all the opportunities you gave me to grow as a researcher and for meeting me where I was as I waded into the world of sociolinguistics. I wouldn’t have done even half of what I did without your mentorship. To AMT, ‘something corny’. To Savi, thank you for running with the native speaker reading group idea that Carrie Ann and I couldn’t entirely get off the ground. So much good work has come of it. To the folks in Weinberg, Rick Lewis, Emily Atkinson, and Mara Bollard, thank you for challenging me to think beyond my linguistic silo. To Mara Bollard specifically, thank you for making me a better teacher and a more well-rounded thinker.

This dissertation would’ve been nothing without the help and guidance of my committee. Thank you Chandan, for approaching me in New York at the LSA in 2019 and talking with me about my work. It was the boost I needed on a nerve-wracking day and the support you’ve given since has been incredible. I look forward to the next Carrom releases. David, you shaped how I thought about perception early on in my time here at Michigan, thank you for everything and onwards to smell and taste! Jon, it’s a shame that COVID had to derail the original version of this dissertation but working with you in neurolinguistics was the jolt I needed when I was ready to walk away from the field. Thanks for being that shot in the arm when I needed it most. To Pam, what can I say that we already haven’t said before. Thank you for everything. It was the privilege of a lifetime to work so closely together for so many years. I look forward to a drink on the back deck in August when this all blows over.

Lastly to my family, thank you for all the support you’ve given me throughout my life. You’ve fostered my curiosity and cared for me through good times and bad. The things I’ve done here and will continue to do in the future are just a reflection of your generosity, kindness, and love. To my wife Tamarae, what can’t you do? Getting a PhD was cool but crossing paths with you was even cooler. I look forward to a lifetime of nights after work where we’re not writing anymore and we can do whatever the hell we want. To Augs, you’re the light of my life, but it really would’ve helped if you slept through the night while I finished this. No worries though, snuggling on the couch together to get you back to sleep was worth the writing delays. It was an almost nightly reminder that my goal is to be the best father I can to you and not just a nerd worrying about speech sounds.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Listeners' expectations and predictions about their interlocutors during the course of speech perception have had a central role in the development of sociophonetic models of speech perception. These models emphasize that listeners use their expectations about a speaker's social identity and their knowledge of the structured phonetic variation associated with those identities to guide perception as listeners perceptually adapt to their interlocutors. In a similar vein, multisensory integration has also been shown to modulate linguistic percepts when listeners are presented with incongruent auditory and visual cues. Under these conditions visual cues eclipse auditory information generating illusory percepts where listeners report hearing what they see. This dissertation explores the intersection of these two literatures and probes whether dialect specific visual signals facilitate socially indexed perceptual adaptation without acoustic reinforcement. Two experiments were conducted using illusory stimuli to assess how Indian English and American English participants shifted their categorization and speech shadowing strategies when listening to model talkers from each dialect.

In both experiments, participants interacted with illusory stimuli that were comprised of an acoustic voiced bilabial stop paired with visual articulations constructed to induce coronal percepts (alveolar or retroflex stops or interdental fricatives) or labial percepts (labiovelar approximants or labiodental fricatives or approximants). Crucially, these stimuli required participants to confront phonological substitutions, mergers, or splits from the other dialect, which they were exposed to through a video that provided experience with the real-world speech patterns of the model talkers. For example, in labial conditions, American English participants were predicted to learn a merger between their labiodental fricative and labiovelar approximant categories that would reflect the labiodental approximant in Indian English, while Indian English participants were predicted to learn to split their labiodental approximant category into labiodental fricatives and labiovelar approximants to reflect the American English contrasts.

In the categorization experiment, participants’ categorization of illusory stimuli before and after exposure suggest that, as predicted, American English participants learned to shift (in a subset of conditions) their categorization strategies, reflecting the merged patterns of the Indian English model talker even in light of multisensory incongruity. In comparison, Indian English participants’ responses only showed within dialect categorization shifts, but in a manner that was rooted in both model talker specific expectations and experience.

In the shadowing experiment, conducted to assess intra-category sensitivity to illusory stimuli, participants produced baseline productions of target words and, after receiving experience with the model talkers, shadowed, or imitated, illusory (audiovisual incongruent) and veridical (audiovisual congruent) stimuli. F2-F3 values were extracted to measure post-exposure shifts in production of coronal stimuli and normalized F2 values were extracted for labial stimuli. Measures were compared across the baseline, veridical, and illusory conditions. Against predictions, results showed that model talker specific shadowing was confined to veridical conditions, where both participant groups imitated congruent audiovisual stimuli across dialect boundaries, a finding that suggests that imitation may require acoustic reinforcement.

This dissertation makes multiple contributions to research on linguistic expectation and perceptual adaptation. Results of the categorization experiment suggest that sociophonetic perception persists in light of multisensory incongruity given the right linguistic experience and broad inter-category measures. Results from shadowing suggest that imitation may need acoustic reinforcement when targeting socially indexed intra-category variation. Taken together, illusions provide a novel path forward for researching the structure of sociophonetic perception and whether these percepts depend on acoustics.

Chapter 1 Introduction

When perceiving speech, listeners interpret the acoustic speech signal as meaningful linguistic forms. Although much remains to be learned about the mechanisms and representations underlying this process (e.g. Fowler & Iskarous 2013), there is broad understanding that speech perception is malleable and dynamic. Listeners' perceptual strategies are sensitive to phonetic (Whalen 1984, McMurray et al. 2002) and social (Foulkes & Docherty 2006) variation, the environment in which perception takes place (Lombard 1911, Brumm & Slabbekoorn 2005), and the congruency or incongruency of multisensory cues (Tiippana, 2014). Perceivers benefit from multiple, co-occurring sources of information, which often provide signal redundancy and aid in the perceptual process. This is in contrast to the historical view that listeners perceive invariants in the input acoustic signal and that coarticulatory, social, or environmental variation should be treated as noise that requires normalization or filtering in service of detecting a strictly linguistic signal (Chomsky & Halle 1968).

For example, in speech perception, providing listeners with meaningful social information about a speaker (e.g. about their gender, age, or ethnicity) has been shown to bias their linguistic decision making and facilitates the resolution of segmental ambiguity through top-down anticipatory processes (Drager 2010). In a similar vein, multisensory integration has also been shown to modulate linguistic percepts. It has long been known that incongruent visual cues can eclipse auditory cues generating illusory percepts (the "McGurk effect"; McGurk & MacDonald 1976) and this same effect has also been shown with incongruent tactile information (Fowler & Dekle 1991, Gick & Derrick 2009). In each of the examples above, as is the case with studying perception broadly, we gain a deeper understanding of our perceptual capacities by strategically ‘breaking’ aspects of the perceptual system, with manipulated experimental stimuli, to examine how our perceptual system copes with diminished or otherwise altered inputs. In turn, we learn not only about the robustness of these systems, but we also gain insight as to how these systems might utilize information from other domains (e.g. social, environmental, or coarticulatory sources) in service of facilitating perception and linguistic interpretation.

At the intersection of the literatures on sociophonetic and multisensory perception lies the broad aim of this dissertation, which is to understand how socially indexed acoustic and visual information affects linguistic processing. Specifically, this dissertation asks whether English-speaking listeners, when exposed to the speech of a speaker from another variety of English that systematically differs in targeted patterns, are more likely to arrive at illusory percepts that are faithful to their own phonology or that reflect the phonology of the other variety. This is investigated via two experiments. The first targets how listeners categorize illusory stimuli from two English speakers and tests whether their categorization rates change over the course of the experiment as a function of their experience with the veridical (audio-visually congruent) speech patterns of these speakers. The second targets how listeners imitate illusory and veridical stimuli and examines whether imitations of illusory stimuli pattern with participant baselines (suggesting a reliance on the listener’s own phonology with little socially motivated anticipatory processes) or veridical imitations (suggesting the use of socially motivated anticipatory processing despite the lack of bottom-up cues due to multisensory incongruity). Because listeners come to these tasks with knowledge about sociologically and physiologically constrained patterns of linguistic variation, illusory percepts provide a unique context in that these percepts are not driven by resolving ambiguities in the acoustic signal; rather, the integration of visual cues during the course of perception drives the subjectively reported auditory percepts of listeners.

1.1 Perception of Multidimensional Speech

1.1.1 Theories of Speech Perception

Linguistic theories of speech perception are broadly divided into two theoretical camps – acoustic theories of speech perception and gestural theories of speech perception. The first sees speech acoustics as the object of speech perception (Ohala 1996, Diehl et al., 2004, Redford & Baese-Berk 2023, for a review), that is, cues or features within the acoustic signal are perceived and interpreted as linguistically meaningful perceptual objects for perceivers. From this perspective, speech acoustics are arguably *the* critical component to speech perception and other sensory information from visual or tactile modalities are secondary reinforcement of the acoustics. Historically, this has led to a long search for the invariant acoustic cues that trigger perception which has largely failed to find acoustic invariance that persists across speech contexts. Rather, what invariance that could exist, given that speech perception actually takes place in the real world, may be between acoustics and mental representations of linguistic categories (Stevens 1989). To complicate matters further, acoustic theories of speech perception also require a mechanism by which acoustic perceptual landmarks can be reinterpreted or translated as motor actions if perception and production are to operate using shared representations – which is theoretically desirable to avoid duplicate representations for the perceptual and production modalities. Despite these historical and theoretical hurdles, adherents to acoustic theories of perception are the majority within linguistics with phonological feature theories relying heavily on acoustic descriptions and acoustic correlates (Chomsky & Halle 1968, Prince & Smolensky 1993/2004, Hayes 2004).

In opposition to acoustic theories of perception, gestural theories of perception argue that speech gestures are the primary object of perception. These gestures, while conceptually tied to physical trajectories of the vocal tract, are thought exist either at an abstract level that reflects the intended linguistic target of a given production (Motor Theory; Liberman & Mattingly, 1985) or uses the physical instantiations of gestures (proximal objects) to perceive the actual (rather than intended) gestural source (distal source) of a speech sound (Direct Realism; Fowler 1986, Fowler 1996, Fowler 2004). In these theories, speech acoustics are a byproduct of the speech process rather than the object of perception. Unlike acoustic theories, gestural theories have been less driven to search for invariance as gestural overlap and coarticulatory information are seen as crucial to the perceptual process[1]. Because coarticulation is lawfully governed, that is, a given linguistic object (e.g. /d/) is still that linguistic object despite acoustic variation across different environments (e.g. /da/, /di/, /du/, /de/, /do/), coarticulation helps reveal to perceivers the foundational aspects of what gives a particular linguistic object its distinctive characteristics. Additionally, by having speech gestures be the object of perception, a "shared currency" (Fowler 2004) is created between perception and production that ensures a single representation that can be used by both perceptual and production modalities. However, one area in which little progress has been made within gestural theories has been in the domain of speaker specific normalization and sociophonetic adaptation.

[1] Fowler (1996) notes that, 'by hypothesis, [acoustic structure in speech] will be found to cause specifiers or invariants' given that the source of these specifiers are phonological gestures of the vocal tract. As such, the search for invariance has not been at the forefront of Direct Realist gestural theories.

1.1.2 Sociophonetic Perception

While variation, including variation introduced by social information for a given speaker, has been viewed historically as noise that was needed to be overcome in service of linguistic interpretation, recent work has instead focused on how listeners' sociophonetic knowledge about the speakers who the listener is hearing can be used to enhance linguistic interpretation during perception (Sumner et. al, 2014; Kleinschmidt, 2018). Within speech perception, this has largely been through the use of exemplar models where acoustic variability is preserved in phonetic traces at lexical level of representation (Goldinger 1998, Pierrehumbert 2002, Johnson 2006). Despite its heterogenous nature, socially indexed phonetic variation is highly structured and is often partitioned along socially relevant hierarchies within a particular language or dialect group. This structure is called indexicality and is the relation between a linguistic sign and a real-world object, within a particular linguistic context, towards which that sign points or indexes. This can take many forms – a sign might be a shibboleth (e.g. coke v. soda v. pop), a phonetic feature perceived as an accent (e.g. a fronted [u] or a raised [æ]), or a syntactic construction from particular regional dialect (e.g. The car needs washed). For example, in the shibboleth example presented in the parenthetical above, an utterance of 'pop' might indirectly point toward the fact that a speaker is from the American Midwest. Crucially, indexes are theorized to contextually refer to objects *indirectly* (Peirce & Hoopes, 1991) and as such are susceptible to reinterpretation across speakers and listeners in a variety of contexts. Formulated this way, indexes can be thought of as social perceptual units that drive the interaction between linguistic content and one's ability to situate themselves socially (Silverstein 2003, Eckert 2008, Drager & Kirtley 2016, Jaeger & Weatherholtz 2016).

Accordingly, because of this link between social and linguistic information afforded by indexicality, patterns of phonetic variation can be also used to aid listeners during social evaluation. While it is unclear whether social and linguistic knowledge is stored together in a single representation as formulated in exemplar models (Pierrehumbert 2002, Johnson 2006) or integrated online during speech processing (Sumner et al. 2014, Kleinschmidt & Jaeger 2015), the effect of one dimension of knowledge, be it linguistic or social, on the other is seen not only in the bottom-up processing of socio-indexical features but, additionally, through anticipatory processes such as listener expectation and prediction (Strand & Johnson 1996, Strand 1999, Hay et al. 2006, Staum-Cassasanto 2008, Hay & Drager 2010, McGowan 2016, Bouavichith et al., 2019, Wade 2022).

In studies probing these processes, participants receive cues about some social index (e.g. gender, race, age, etc.) attributed to a speaker and are asked to make a linguistic decision about what they perceive the speaker to be saying. A priori, participants are expected to come to these tasks with knowledge of the meaningful social indexes of their speech community via their linguistic experience; experimental stimuli are then created that contain the phonetic and social cues hypothesized to activate these indexes. As a result, when participants are presented with stimuli that contain ambiguous linguistic cues, ambiguities that are artificially controlled in the lab but in many cases are not unlike the ambiguities present in conversational interactions, listeners rely on the auditorily or visually cued social indexes to aid in disambiguating the speech signal. For example, Strand & Johnson (1996), Munson (2011), and Bouavichith et al. (2019) have all shown that listeners presented with a visual gender cue (e.g.

an image of a male or female face) paired with an acoustic fricative stimulus with a center of gravity between /s/ and /ʃ/, show a shift of their perceptual boundaries for /s/ - /ʃ/. The nature of this shift suggests that listeners rely on their experiential knowledge of speaker physiology or sociolinguistic patterning to more effectively parse the speech signal. Consequently, the same auditorily ambiguous /s/ - /ʃ/ token is categorized as /s/ by listeners when the visual cue is male and as /ʃ/ when the cue is female. This aligns with the acoustic consequences of vocal tract length variance where, due either to gendered performance or speaker physiology, shorter vocal tracts (associated with canonically feminine speakers) show a distribution of higher frequency sounds and longer vocal tracts (associated with canonically masculine speakers) show a distribution of lower frequency sounds. As such, when participants are presented with ambiguous sibilants that lie in the overlap of distributions for canonically feminine /ʃ/ and canonically masculine /s/ (the orange area in Figure 1.1) it appears that the visually communicated social information is what drives listener linguistic categorization strategies.

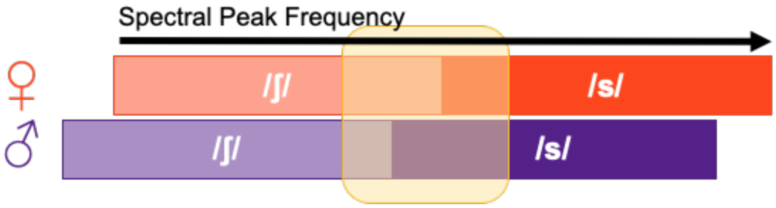

Figure 1.1: Spectral peak frequency distributions for canonically feminine and masculine presenting speakers. Bouavichith et al. (2019) sampled from the yellow square where /ʃ/ as produced by canonically feminine speaker and /s/ as produced by canonically masculine speakers overlap.

In Bouavichith et al. (2019) the authors also tested for effects in the reverse condition: whether presenting visual linguistic information (e.g. black and white drawings of a *sack* or *shack*) influences listener judgments of perceived speaker gender. They reported a bidirectional effect: listeners categorized gender-ambiguous voices as feminine when presented with a visual cue for a lexical item beginning with an /ʃ/ onset and masculine when presented with a visual cue for a word beginning with an /s/ onset. This finding supports what has been theorized to be a complex and interactive relationship between linguistic and social information and how listeners can use the indexical relationships between these information streams to arrive at percepts that are both linguistically and socially meaningful. Specifically, Bouavichith et al. (2019) provides experimental evidence that suggests that listeners use the multidimensionality of the speech stream (i.e. the idea that the speech stream is comprised of a social dimension with information about a speaker or groups of speakers and a linguistic dimension about the language content of an utterance), and that when one dimension of the speech stream contains ambiguity, that listeners can resolve it by integrating sufficiently unambiguous information from the other dimension in conjunction with their indexical knowledge about the relation between the two dimensions.

Socially modulated linguistic categorization is not only constrained to gender though; providing participants with socially indexed visual stimuli also modulates linguistic responses to stimuli that are socially indexed for speaker age (Hay et al. 2006), race (Staum-Cassasanto 2008), regional or national dialects (Niedzielski 1999, Hay & Drager 2010), and the perceived foreign accentedness of a speaker's English (McGowan 2016). Additionally, in online perceptual tasks with acoustic stimuli, listeners have been shown to update their perceptual strategies when presented with talker-specific patterns of structured variation over the course of an experiment consistent with normalizing for speaker-specific patterns of variability (Bertelson et al. 2003, Norris, McQueen & Cutler 2003, Kraljic & Samuel 2006, Trude & Brown-Schmidt 2012, Reinisch & Mitterer 2016). This effect is achieved by exposing participants to acoustic information (e.g. a recording where the formant transitions of a stop are ambiguous between /t/ and /d/ or a recording with vowels of a particular accent type, Kraljic & Samuel 2006) then, in a training phase, participants are provided disambiguating cues either lexically or visually. Lastly, participants undergo a test phase where categorical judgments to the acoustic stimuli are collected which often exhibit effects of a perceptual recalibration based on their experience within the experiment.

Taken together, studies like these again reflect the complexity of social-linguistic interactions and detail the susceptibility of linguistic percepts to shift when a signal shows ambiguity or instability along one dimension that can be resolved through signal definition or stability along another dimension. I target this relationship between social and linguistic dimensions within the speech signal in this dissertation by probing the degree to which these socially modulated adaptation effects persist in illusory conditions where the acoustically indexed perceptual cues that typically trigger a social index have to be imputed by listeners while resolving a multisensory incongruency.

1.1.3 Illusory Perception and the McGurk Effect

Visual modulation of auditory speech percepts is not only limited to the realm of social meaning making. Research investigating the role of visual and tactile perceptual integration during the course of speech perception, famously via the McGurk effect, has also revealed how the reported linguistic percepts of listeners can be modulated by non-acoustic perceptual information. This illusory phenomenon occurs when listeners receive incongruent audio and visual (McGurk & McDonald 1976) or incongruent audio and tactile information (Gick & Derrick 2009, Ito et al. 2009). Traditionally, in trials where acoustic and visual cues are mismatched, such as Acoustic /ba/ paired with Visual /ga/ (AbVg), visual information appears to eclipse auditory information which in turn drives the reported percept of listeners: typically reported as /da/ in the case of AbVg[2]. In a trial with an incongruency such as this one, it is theorized that a listener, upon hearing a bilabial stop release but seeing no visual evidence of a lip closure, either weights or prioritizes the visual signal and arrives at a percept for a more posterior sound. When played the same stimulus with no visual cues (e.g., viewing a screen with no corresponding image or closing one's eyes) participants report a percept reflecting the acoustic content of the stimulus, in this case /ba/.

In a similar vein to the theories outlined in § 1.1.1, theoretical explanations regarding the perceptual primitives involved in experiencing or perceiving an illusory McGurk effect have also been split into theorizing about whether the object of perception is acoustic or gestural in nature. Here these differences are somewhat more tangible given that the effect depends on one modality eclipsing the other. In general multisensory theories (Shams 2011), acoustic information is prioritized as the object of perception and the visual stream is seen as a support system intended to reinforce what the ears are hearing. Listeners learn associations between the auditory and visual streams and, in the case of McGurk stimuli, they rely on these learned associations to evaluate and re-evaluate mismatches between the acoustic and visual streams. In this way, vision can be prioritized or weighted to augment confusable acoustic information. In gesturalist approaches, often called supramodal approaches (Fowler 2004, Rosenblum et al. 2016), the shared information between vision and audition are not seen as learned associations but a form of "common currency" about a single action of the vocal tract

[2] Tiippana (2014) quotes McGurk & MacDonald (1976) who note that, "…lip movements for [ga] are often misread as [da]". Tiippana goes on to point out that while McGurk & MacDonald didn't provide any measure of speech reading performance in their paper, in her own work (Tiippana et al., 2004) she finds that confusability between visual [ga] and [da] contributes to identifications of AbVg stimuli as perceptual [da]. In any case, participants experiencing the illusion report 'hearing' the visual stimulus, as opposed to the acoustic stimulus.

(Goldstein & Fowler 2003 p.174). In this way, the object of perception doesn't need modality specific information channels. Rather, each unique configuration of the vocal tract imparts information across a string of modalities about a single perceptual event. It is the event listeners perceive not its physical fingerprints.

Broadly, visual and auditory streams appear to be treated similarly when perceiving speech (Rosenblum 2005, Rosenblum et al 2017). Not only does congruent visual information facilitate auditory speech perception (e.g. in noise; Sumby & Pollack, 1954, Bernstein et al. 2004), but also, listener experience across one modality (e.g. lipreading) appears to facilitate perception in the other modality (e.g. auditory only identification, Rosenblum et al. 2007). Likewise, visually driven McGurk stimuli appear to activate similar neural sources as acoustic only stimuli (Sams et al 1991, Colin et al. 2002, Saint-Amour et al. 2007). While these studies have highlighted similar neural sources between McGurk percepts and acoustic percepts, others (Beauchamp et al. 2010) have shown differential effects of transcranial magnetic stimulation that could suggest separate neural sources for McGurk and veridical audiovisual percepts. This has led to an active and open debate about the perceptual equivalence of illusory and veridical perception and whether these forms of perception have the same computational bases (Alsius et al., 2018; Hickok et al., 2018; Rosenblum, 2019) or whether illusory perception requires additional neuronal processing to arrive at percepts.

Questions have also been raised about the ubiquity of the effect in listeners. McGurk percepts have been shown to be highly variable across listeners and it is sometimes the case that particular listeners never report perceiving them at all while others readily perceive the effect (Nath & Beauchamp, 2012). In a similar vein, McGurk effects are heavily dependent on the particular stimuli used by researchers (Basu Mallick et al., 2015) as similar but not the same configurations vary in their magnitude of effect across studies. Phenomenologically, it is known that McGurk percepts are often ‘fuzzier’ than audiovisual congruent stimuli (Alsius et al., 2018) and are susceptible to enhancement or diminishment by varying the timing of stimuli congruency (Munhall et al., 1996; vanWassenhove et al., 2007), familiarity (Walker et al., 1995) and orientation (Eskelund et al., 2015) of the stimulus face, congruency of the visual and acoustic gender of the stimuli (Green et al., 1991), and the native language of participants (Sekiyama & Tohkura, 1991; Hardison, 1999; Sekiyama & Burnham, 2008; Burnham & Dodd, 2017), all of which leave the strength of the effect open to influence by other cognitive mechanisms.

The vowels surrounding an illusory consonant can also condition the consonantal identity of the illusory percept. Green & Kuhl (1991), Burnham (1998), and Shigeno (2002) have found that audio-visual incongruencies of the form AbVg led English-speaking participants to report more /ð/ responses when the illusory consonant is surrounded by [a] and more /d/ responses when surrounded by [i]. Burnham (1998), Sekiyama & Burnham (2008) and Burnham & Dodd (2017) have used this contextual asymmetry to investigate how differences in the phonemic inventories of Australian English (which includes /ð/) and Japanese (which does not) account for reported McGurk percepts by English- and Japanese-speaking participants when listening to English and Japanese speakers. These studies found that the response asymmetry reported above is maintained by all participants in trials with an English stimulus speaker but, when a Japanese speaker is presented, both Australian English and Japanese participants report fewer [ð] percepts. However, both sets of listeners still showed the vowel conditioned rates of [d] responses across both vowel contexts. The authors interpret these results as evidence of early phonetic low-level processing in McGurk illusions (in the case of the vowel-conditioned asymmetries) that is left open to later stage language-specific or cross-cultural effects (the absence of [ð] responses for the Japanese speaker). This leaves open the potential that cultural knowledge on the part of Australian listeners, and phonotactic knowledge on the part of Japanese listeners about the lack of an /ð/ phoneme in Japanese, were the drivers of [ð] response absence.

However, these two forms of knowledge, cultural (or sociolinguistic) and phonotactic are potentially at odds with one another at a functional level of explanation (Marr; 1982) despite resulting in the same measurable experimental outcomes. In the case of Australian English speakers, it seems as though knowledge of the speaker's grammar, in this case Japanese, is the lens through which the listener hones their percepts at the late stage of perception. In the case of the Japanese listener, however, there is the potential that it's the listener's own grammar, again Japanese and in this case shared with the speaker, through which the listener hones their percepts. Teasing apart these possibilities is one specific aim of this dissertation. And, rather than using languages that differ in phonemic inventory or phonotactic knowledge, I use two dialects which differ in phonological form but are intelligible to one another allowing for contextual disambiguation of accents and index creation. By allowing for mutual intelligibility, we can better understand the output of categorization of illusory stimuli not strictly as an output of *what* was said but *how* it was said; a reflection of the speaker's grammar or the listener's grammar. To understand the question of *what* is being said, I use a phoneme identification paradigm, as is traditional in illusory perception research. But, for the more fine grained question of *how* an illusion is said, I use a speech shadowing paradigm, which is arguably better suited for examining the (subcategorical) phonetic characteristics of what participants are perceiving.

1.2 Speech Shadowing

Humans are natural language imitators, and it is widely understood and accepted that speakers converge towards one another's speech productions in both discourse and lab settings (Dufour & Nguyen, 2013; Pardo, 2013). While this imitation is undisputed, much of the shadowing literature has centered around debates about whether the motivations for imitation are linguistic or social (Mitterer & Müsseler, 2013; Nielsen, 2011), whether the representational content that is accessed is gestural or auditory (Honorof et al., 2011; Mitterer & Ernestus, 2008; Shockley et al., 2004), and whether this representational content is lexical, phonological, and/or phonetic in nature (Fowler, 2003; Goldinger, 1998; Honorof et al., 2011; Mitterer & Ernestus, 2008; Nielsen, 2011, Kwon 2019). There is also a small literature that studies shadowing in illusory contexts that is particularly relevant for this dissertation (Gentilucci & Cattaneo, 2005). Each of these literatures provides helpful insights to the work proposed here and their contributions are outlined below.

Speech shadowing as a methodology received its most recent revival from work by Goldinger (1998) who argued that word representations, perception, and production were episodic in nature – that is, the lexicon is comprised of detailed acoustic episodes as opposed to abstract phonological representations. Goldinger (1998) had participants listen to shadowed speech and perform AXB discrimination tasks in which participants judge the similarity of X to A vs. B; in this case, participants judged whether a shadower's baseline (A or B) or shadowed (B or A) stimulus was more similar to the speech of the model talker (X stimulus) being imitated by the shadower. Across multiple experiments, in addition to finding effects of frequency and repetition, Goldinger found that participants in the AXB tasks reliably perceived participants' shadowed speech as being more like the speech of the model talker, suggesting that, in spontaneous imitation, aspects of a model talker's production patterns influence the productions of a shadower. Evidence of this model talker similarity effect in AXB judgements is what Goldinger used to argue that shadowed productions must utilize perceptual episodic experiences retained in the memory of the shadower and which are consequently perceived by listeners.

While AXB tasks provide global judgments about the perceptual similarity between two stimuli, they do not allow us to know which specific phonetic dimensions of shadowed stimuli are being imitated by shadowers. Fowler et al. (2003), using vowel-consonant-vowel (VCV) syllables, and Shockley et al. (2004), using bi-syllabic English words, investigated at how particular phonetic dimensions (e.g. voice onset time; VOT) are imitated when participants shadow model speech. In these studies, the authors had participants shadow model speech where VOT had been artificially extended in a subset of test conditions. It's worth noting that in these extended conditions, VOT did not encode a phonological contrast; rather, long lag English stops just became longer. In both the work of Fowler et al. (2003) and Shockley et al. (2004) shadowers also extended their VOT between the baseline and shadowing conditions suggesting that shadowers perceived the variable of interest and imitated it in their own productions. The authors used these results to put forward the hypothesis that shadowers don't necessarily have to be using episodic memory as argued by Goldinger (1998), but that they may just be tracking gestures in real time during shadowing and imitation. Nielsen (2011) extended this line of experimentation by showing that shadowing appears to be sensitive to phonological boundaries and categoricity. Rather than only extending VOT (Fowler et al. 2003; Shockley et al. 2004) Nielsen also reduced VOT so that it approached the perceptual boundary between /p/ and /b/ in English. Nielsen found an asymmetry in imitation when it came to a reduction of VOT. In conditions with reduced VOT, shadowers did not imitate the model talker, which Nielsen argued was evidence that imitation is not only sensitive to phonological boundaries but selective as well. The fact that shadowing appears to be sensitive to phonological boundaries is critical for this dissertation because it is the phonological knowledge of participants that is being leveraged when they are presented illusory stimuli during the shadowing experiment.

However, one crucial question that lingers given Nielsen's findings is how cross speaker-shadower phonological variation is captured with shadowing experiments. Recent findings by Schertz & Paquette-Smith (2023) have found that listeners can in fact converge towards shortened VOT in a task where imitation is explicit. In their work, participants were explicitly instructed to imitate the speaker that they heard in addition to completing a discrimination task. Participants were able to not only imitate the shortened VOT conditions, but also showed greater discrimination of shortened VOT conditions. The authors argue that the correlation between greater discrimination and shortened VOT convergence in their results suggests that there is no general constraint against shadowing at a phonological boundary but that perceptual salience can facilitate within category imitation even for non-canonical category exemplars. Mitterer & Ernestus (2008), Honorof et al. (2011), and Mitterer & Müsseler (2013) also provide insight into how shadowers handle imitating phonological forms that map onto multiple phonetic realizations. Work led by Mitterer has largely centered around understanding how shadowers handle gestural mismatches between two phonetic realizations of a single phonological category. Mitterer & Ernestus (2008) looked at shadowing differences between [ʀ] and [r] which vary on a speaker by speaker basis in Dutch and Mitterer & Müsseler (2013) investigated shadowing differences between German word initial [st-] and [ʃt-], which are regionally distributed, and word final [-ik]/[-iç], which freely vary. In both of these studies, the authors found no reaction time differences in the onset of shadowing, which they interpret as showing that shadowing takes place at the phonological level given that one might expect the gestural mismatch between a shadower's preferred phonetic production and the model talker's production to induce a delay in shadowing (e.g. shadower [ʀ] and model talker [r]). However, Mitterer and colleagues (2008, 2013) also note that shadowers tend to stick to their preferred phonological pronunciation (baseline) unless the phonological variation they are shadowing is either socially salient or participants are instructed to "correct" or "imitate" a model talkers production[3]. This lends further evidence to Nielsen's claim that, while imitation naturally occurs, it may be a more selective process than an automatic one.

Honorof et al. (2011) also looked at how phonological variation is handled in shadowing but when conditioned by syllable position as opposed to conditioning by dialect specific or free variation patterns. Honorof and colleagues used VCV non-word stimuli to understand how /l/, which differs in its realization as a function of whether it is syllabified as an onset [l] or a coda [ɫ], is shadowed by American English listeners. They found that shadowers were able to reliably perceive and imitate the two different types of American English /l/ despite the ambiguity in syllabification. Further, these imitations were assessed using articulatory sensors to show that shadowers imitated the approximate articulatory gestures of the model talker. Unlike the work led by Mitterer, the authors note that participants freely imitated the model talkers, potentially suggesting a difference for how cross-dialect (e.g. /r/ → [r], [ʀ] in Mitterer & Ernestus, 2008) and within-dialect (e.g. /l/ → [ɫ]/_# in Honorof et al., 2011) phonological motivations condition the production strategies that shadowers use.

[3] Dufour & Nguyen (2013) also report on the role that instructions play in determining the magnitude of a shadowing effect (e.g. "imitate what you hear" v. "repeat what you hear").

Lastly, there is the question of how speakers shadow illusory stimuli, which has been examined by Gentilucci & Cattaneo (2005). In their work, the authors had two groups of Italian participants shadow congruent and incongruent VCV stimuli (e.g. ‘McGurk’ AbVg, ‘Inverse McGurk’ AgVb) while measuring participant lip aperture and the F1/F2 acoustic characteristics of shadowed productions. In the traditional McGurk (AbVg) condition, the authors analyzed productions from participants who reported hearing [aba] (n = 21) or [ada] (n = 8). In the latter case of illusory fusion (i.e. a reported [ada] percept) the authors found that shadowed responses to incongruent stimuli did not differ statistically from congruent shadowed productions, suggesting that cases of illusory fusion contained sufficient perceptual clarity to mirror veridical shadowing from the congruent (AdVd) conditions. For the Inverse McGurk condition, the authors reported differences in F2 between congruent (AgVg) and incongruent (AgVb) stimuli when participants fail to have an illusion, suggesting that aspects of the visually presented [aga] syllable still influence the shadowed [aba] productions. This not only serves as further evidence of the ability of vision to eclipse audition in illusory stimuli but also suggests a similarity of sorts between ‘successful’ illusions and veridical perception such that shadowing lends itself as an appropriate methodology for assessing how much within category variation is imitated when a speaker is experiencing a McGurk effect.

Taken together, this literature not only establishes shadowing as a valuable experimental tool for understanding how perceived within category variation is imitated in the production patterns of speakers, but it also highlights that shadowed productions are crucially sensitive to phonological, sociolinguistic, and multisensory perceptual dimensions. Further, the fact that there appears to be a trade-off between the preferred production patterns of the participant and the phonetic targets of the model speaker motivates the use of mutually intelligible dialects as they contain ‘marked’ productions of shared lexical items between dialects. This leaves open the question that I pursue via shadowing: how do listeners with different phonological grammars – and, as a result, different perceptual boundaries – interpret the same illusory manipulations given the illusion’s lack of socially relevant acoustic inputs?

1.3 The Role of Illusions

Illusions play a crucial role in this project and the significance of investigating this project’s specific illusory context goes beyond simply delineating the nature of illusory percepts. The primary goal of this project is to reveal the degree to which socially indexed listener expectations about what talkers are saying, expectations borne out from listener experience, are dependent on acoustic signals. The effects that arise from socially indexed anticipatory perceptual processes have primarily been situated in the literature within descriptions and analyses that rely on the retention of representational distributions of acoustic cues or phonetic exemplars held in memory or the lexicon (Goldinger 1998, Pierrehumbert 2002, Johnson 2006, Kleinschmidt 2018); largely acoustic domains. However, in McGurk illusions, information within the visual signal eclipses information within the auditory signal, compromising the auditory signal’s utility to the listener in perception. Given that articulatory speech gestures causally generate the multisensory speech signals transmitted around us, it stands to reason that the visual signals generated from these gestures are not only specified with linguistic information that is perceptually significant, as evidenced when perceivers experience a McGurk illusion, but also are causally specified for socially relevant information that can be used for experiential perceptual learning. The aim of this dissertation is to use illusions as a strategic tool to probe the degree to which these linguistically specified visual signals can facilitate the socially indexed expectational processes reported elsewhere in auditory accounts and to assess whether the outputs of those processes are perceptually equivalent to perceiving veridical signals, those in which audio and visual information are congruent.

To address this question, I examine how phonetic differences between two mutually intelligible English dialects, Indian English and American English, affect behavioral responses of participants as they gain experience with the phonological patterns of speakers from each group. The hypothesis that guides this work is that listeners utilize all available meaningful information from an incoming signal and that illusory percepts reflect a unique experimental case that targets what listeners impute on conditions that are intentionally constructed to be imperceptible from the union of their incongruent component parts. By design, illusions do not allow perceivers to experience a unified percept that is faithfully representative of all its constituent parts. Rather, what we perceive is a computable solution to a conceivable but non-computable problem. Put another way, illusions reveal foundational aspects of the perceptual process that have developed across a listener's lifetime, and which persist despite stimulus incongruence (e.g. perceiving what you see rather than what is presented aurally).

1.4 Indian English & American English

Central to this project are dialect-specific differences between Indian English and American English within the labial and alveolar-to-velar articulatory regions, regions that are susceptible to McGurk effects. By targeting these particular articulatory regions, this work attempts to use multisensory incongruity as the locus of speech ambiguity to understand whether listeners will incorporate the dialect-specific patterns of substitution outlined in Table 1.1 to generate percepts that are sociolinguistically meaningful.

Lexical Item	Mainstream US English	Indian English
the	[ð]	[d]
brea**th**e	[ð]	[d]
whe**th**er	[w], [ð]	[ʋ], [d]
wa**v**es	[w], [v]	[ʋ], [ʋ]
vo**w**els	[v], [w]	[ʋ], [ʋ]

Table 1.1: Pronunciation differences between American English and Indian English. Orthographic bolding corresponds to International Phonetic Alphabet transcriptions.

Traditionally, Indian English has been framed as a case of first language interference (e.g., Indo-Aryan or Dravidian) on English or Received Pronunciation (RP) phonological targets. However, contemporary phonological descriptions have argued that Indian English has its own nativized phonological targets that are shared across the sub-continent (Gargesh 2008, Sailaja 2012, Wiltshire 2020) and that listeners of other Englishes perceive these patterns in such a way that Indian English speakers can be uniquely grouped by listeners (McCullough & Clopper 2016). As such, throughout this work I will be treating Indian English as a phonologized dialect in the same way one might write about New Zealand English or African American Vernacular English. Certainly, there are raciolinguistic dimensions to how Indian English is perceived both socially and linguistically by listeners within India and across the world (Sonntag 2009, Rosa 2016, Rosa & Flores 2017). What I am asserting here is that I will not be conceptualizing Indian English as some form of English as filtered through Hindi, Bangla, Marathi, Telugu, Tamil, or any other number of Indo-Aryan,Dravidian, or Tibeto-Burman languages found in India. Rather, I will treat Indian English is a phonologized dialect with its own unique phonetic targets that are mutually intelligible with other English dialects. In this way, I see the differences between Indian English and American English less as a difference of type, which might presuppose that the experimental tasks proposed in Chapters 2 and 3 are about non-native perception, and more as a difference of degree about how speakers of the same language adapt to dialectal differences.

As seen in Table 1.2 and Table 1.3 below, American English and Indian English share a large proportion of their consonantal inventories. Two specific differences between the phonologies of Indian English and American English are the foci of this dissertation. The first is the realization of the interdental fricative [ð] in American English as [d] in Indian English. This is a somewhat common substitution across world Englishes (e.g. African American Vernacular English, Michigan Upper Peninsula English, Cajun English) given the rarity of [ð] across the worlds languages and its susceptibility to fortition (Zhao 2010). The second phonological difference is the merger of the distinction in American English between [v] and [w] to [ʋ] in Indian English. Indian English speakers are reported to collapse these two sounds from American English into a single labio-dental approximant that shows evidence of both frication and lip rounding (Fuchs, 2019). While speakers of American English use these properties individually to encode contrast—[w] has lip rounding but no frication and [v] has frication but no lip rounding—[ʋ], found in Indian English, not only shows evidence of both properties but also is interchangeable with orthographic ‘w’ and ‘v’[4].

[4] Gargesh (2008) reports that speakers of Odia and Bangla produce ‘/v/ as [b^h] in words like *never* [neb^{h}ər] in Indian English’. There were two participants, one in each of the experiments described in Chapters 2 and 3, who either reported speaking Bangla or who reported growing up in Indian states where Odia and Bangla are widely spoken; see sections 2.1.1 and 3.1.1 for more detail.

	Bilabial	Labio- dental	Dental	Alveolar	Post- Alveolar	Palatal	Velar	Glottal
Plosive	p b			t d			k g	
Affricate					tʃ dʒ			
Nasal	m			n			ŋ	
Fricative		f v	θ ð	s z	ʃ ʒ			h
Approximant				ɹ		j	w	
Lateral Approximant				l				

Table 1.2: American English consonant inventory from Hillenbrand (2003). Of note are the voiced dental fricative [ð], the voiced labio-dental fricative [v] and the voiced velar approximant [w].

	Bilabial	Labio- dental	Dental	Alveolar	Post- Alveolar	Retroflex	Palatal	Velar	Glottal
Plosive	p (p^{h}) b		t (t^{h}) d			ʈ (ʈ^{h}) ɖ		k (k^{h}) g	
Affricate					tʃ (tʃ^{h}) dʒ				
Nasal	m		n					ŋ	
Fricative		f		s z	ʃ				h
Approximant	ʋ/w			r			j		
Lateral Approximant				l		(ɭ)			

Table 1.3: General Indian English consonant inventory from Wiltshire (2020) and CIEFL (1972). Of note are the voiced dental stop [d] and the voiced bilabial or labio-dental approximant [ʋ].

Each of these phonetic realizations is within articulatory regions that are susceptible to the McGurk effect and also require listeners from each dialect to learn substitution patterns for speakers outside of their own dialects (Table 1.1). In the case of consonants in the dental-alveolar region, this requires each dialect group, Indian English and American English, to learn a new a new allophonic relation. For American English participants, this is the equivalent of learning a fortition rule for the Indian English speaker (e.g. /ð/ → [d]; Figure 1.2a). For Indian English participants, this is the equivalent of learning a spirantization rule for the American English speaker (e.g. /d/ → [ð]; Figure 1.2c). While these rules have different phonetic consequences and likely also differ in their phonological naturalness, they are functionally equivalent; both add a new allophone to a phoneme in the listeners' inventory.[5]

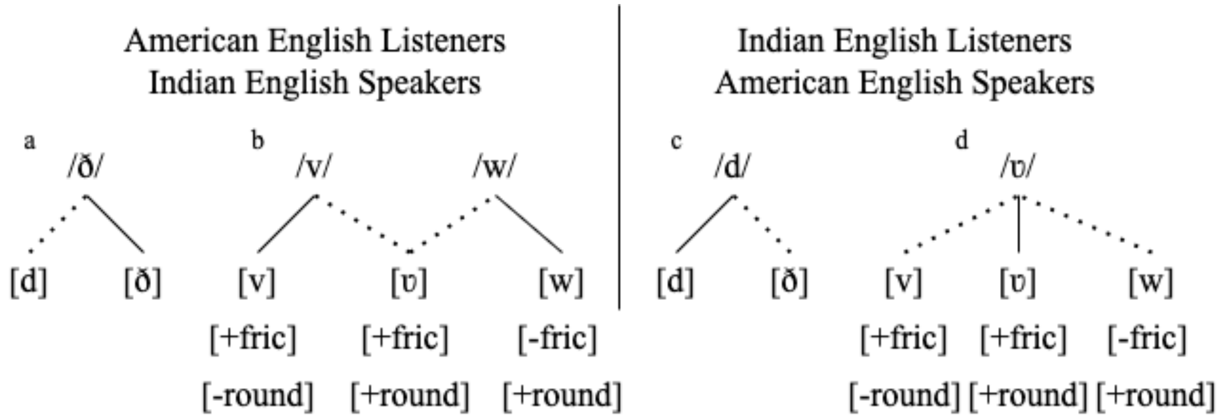

Figure 1.2: Phonological learning required of American and Indian English listeners across dialects. Solid lines represent within dialect allophonic relations. Dotted lines represent learned cross dialect allophonic relations. A) American English participants must learn that underlying /ð/ maps onto [d] in addition to [ð]. B) American English participants must learn that /v/ and /w/ as cued by lip rounding and frication are merged into [ʋ]. C) Indian English participants must learn that underlying /d/ maps onto [ð] in addition to [d]. D) Indian English participants must learn that lip rounding and frication are distinctive features of [v] and [w] in American English.

In the case of labials, participants again learn new allophonic relations but this results in American English listeners learning a merger, and Indian English listeners learning a split. For American English participants, they must learn that frication, a feature associated with [v], and lip rounding, a feature associated with [w], are not distinctive for the relevant sounds in Indian

[5] While I have largely situated the differences between American and Indian English within the gestural articulation and gradient perceptual learning literatures, I use binary features in Figure 1.2 similar to what is found in the Sound Pattern of English (Chomsky & Halle, 1968) for ease of explanation. It's worth noting here that this formulation isn't substantively different from gestural and gradient frameworks, but rather provides a clear visual picture of what is necessary for learning the dialectal differences between American English and Indian English listeners.

English (Figure 1.2b). Rather, Indian English [ʋ] exhibits both frication and lip rounding. For Indian English listeners, they must learn frication and lip rounding are distinctive in American English and, rather than having a single phoneme /ʋ/ with a transparent allophone [ʋ], they must utilize these distinctive features to map /ʋ/ onto /v/ and /w/ (Figure 1.2d).

I target these learning realities through two experiments: one where participants from each dialect (American and Indian English) are asked to categorize illusory stimuli from talkers within and across the participants' dialect and the other where participants from each dialect (American and Indian English) are asked to shadow, or imitate, illusory and veridical stimuli produced by talkers from within and across the participants' dialect. In each experiment, participants complete an experimental task before becoming familiar with the talkers and then again after receiving experience with both the American English and Indian English talkers. Over the course of both the categorization and shadowing experiments, participants are expected to adapt their perceptual strategies to account for each talker's visually cued production differences especially after intervention periods where they are exposed to the veridical speech patterns of talkers from both dialects. While the descriptions in Tables 1.2 and 1.3 are not exhaustive, they are what is minimally required to set the stage for the deeper discussion of specific predictions in the methodology sections of Chapter 2 (Categorization Experiment) and Chapter 3 (Shadowing Experiment).

The remaining chapters of the dissertation are as follows: Chapter 2 presents the methodology, predictions, and results of the categorization experiment. Chapter 3 presents the methodology, predictions, and results of the imitation experiment. Chapter 4 discusses the findings of these two experiments in relation to one another as well as in relation to the literature summarized in this chapter and offers future directions of research.

Chapter 2 Categorizing Illusory Percepts

This study investigates the influence of sociophonetic experience on the categorization of illusory speech percepts by speakers of American English and Indian English. As mentioned in §1.4, American and Indian English differ from one another in their consonantal inventories, importantly for this study, in consonants (e.g. /d/, /ð/, /v/, /w/, /ʋ/) which are susceptible to the McGurk effect. Previous studies (Kraljic & Samuel 2006, Bradlow & Bent 2008) have found that listeners can use socially indexed phonological information to shift their listening strategies as they gain more experience with a speaker. Often, this is achieved through a perceptual adaptation task where participants perform the same experimental task before and after an intervention phase where they gain experience with the accented speech of a model talker or talkers. This accented speech is often created through experimental manipulations that are not unlike patterns seen in the natural speech of talkers.

Illusory percepts present a special case where, rather than controlling the degree to which a token triggers particular sociolinguistic indexes along a continuous acoustic dimension, one can instead acoustically control the *type* of consonant (e.g. Acoustic: [b]) a participant is exposed to and assess the degree to which the listener imputes qualities from a different visually cued consonant (e.g. Visual: [v]) type on what they ultimately perceive (e.g. /v/). As noted in § 1.3, the aim of this dissertation is not to uncover something about how illusory percepts are generated or implemented. Rather the aim is to use illusions as a strategic tool to probe whether linguistically specified visual signals facilitate socially indexed perceptual adaptation without acoustic reinforcement. As such, only a low-level characterization of what goes on in illusory perception is necessary. At their most primitive level, illusory percepts rely on a clear acoustic signal, a clear visual signal, and a phonologically relevant interpretation of the percept that is constructed from those two signals. Crucially, the acoustic signal, while itself unambiguous, can be misinterpreted in the presence of a conflicting visual signal.

In this way, one can probe how differences in phonological knowledge changes the interpretation of the percept constructed from the confusable elements of the acoustic signal (e.g. the broadband burst of [b]) and a clear visual signal (e.g. the labio-dental articulation of [v]). In the case of American English listeners, one might expect that they perceive something akin to /v/ given that the acoustic burst of [b] can be “integrated”, “re-mapped” or “misinterpreted” as frication associated with [v][6]. In the case of Indian English listeners, one might expect them to perceive something between /v/ and /w/ given the near merger of those two phones in Indian English (Gargesh 2008, Fuchs 2019, Wiltshire 2020) and the fact that [ʋ] shows acoustic evidence of frication that can be illusorily perceived from the burst of [b]. This is the relationship that I hope to experimentally probe in this experiment by providing different groups of listeners with illusory stimuli as spoken by speakers from different dialect groups. For this experiment I am explicitly interested in two research questions: 1) Does the proportion of illusions experienced by participants change after gaining experience with the accent of the model talker and 2) Do illusory categorization rates change after participants gain experience with the accent of the model talkers? In the sections below, I discuss predictions for each of the

[6] I am purposely staying agnostic here as to whether what listeners perceive during illusory perception is vision to the exclusion of audition, an integrated percept from the two modalities, or something in between. While this might limit explanatory power for where we think indexical information comes into play in percept formation writ large, in this experiment I simply want to understand the kinds of indexical information available to listeners at the end point of the perception process.

listener groups (Indian English & American English) and how intervention might shift these illusory percepts once listeners have experience with the real accents of the model talkers.

In the following experiment, American English and Indian English speakers watched audiovisual stimuli originally produced by two model talkers, a speaker of Indian English and a speaker of American English. Stimuli were edited to elicit both illusory percepts and veridical percepts. Participants were asked to categorize (according to specified phonemic categories) what they thought the model talker said both before and after an intervention phase during which listeners gained experience with the model talkers' accented speech. Rates of illusory perception as well as specific types of illusory percepts before and after intervention are compared. The basic hypothesis that I pursue is that listeners will use all available meaningful information from an incoming signal, and as such, listeners will experience percepts that reflect their experience with a talker up to the point of performing the categorization task. Thus, I expect that participants will respond with patterns reflecting their own phonology before intervention and reflecting, at least to some degree, the phonology of the model talkers after intervention. The rest of this chapter describes the details of the methodology (§ 2.1), predictions (§ 2.2), results (§ 2.3) and discussion (§ 2.4) for this experiment.

2.1 Methods

2.1.1 Participants

One hundred speakers of American English and 100 speakers of Indian English were recruited via Amazon Mechanical Turk to participate in the study. Postings for the experiment were uploaded to Amazon Mechanical Turk and IP addresses were restricted to the United States for American English recruitment and India for Indian English recruitment. Participants self-identified as English speakers who had learned English no later than between the grades K-12 in the country where they were accessing the experiment[7]. Indian English participants reported speaking one or more of Tamil, Hindi, Malayalam, Marathi, Gujarati, Urdu, or Telugu in addition to English either at home as a child or currently at home. One Indian English participant reported living in West Bengal and knowing Bangla as 'another language they speak' which is a language variety where speakers are known to have [b] as a realization of /ʋ/. This participant also reported that they grew up speaking Hindi and English at home and as such were included in the analysis. In addition to monolingual American English participants, American English participants reported speaking Spanish, Chinese, French, Italian, German, Marathi, Tamil, or Hindi in addition to English. American English participants (4) who reported speaking Tamil, Hindi, or Marathi reported that these were 'other languages' they had learned. This contrasts with the language they reported speaking at home as a child (English) or that they currently spoke at home (English).

Participants who experienced illusory percepts at a rate of chance (33%) or higher in any given experimental block (described in § 2.1.3.1) were included for analysis of that block. This results in some participants appearing in all the statistical analyses while some only appear in a subset of blocks where they experienced illusions at a rate greater than chance.[8] No

[7] This inclusion criterion is intended to capture a shared language experience of being taught a standardized language variety rather than serving as a proxy for age of English acquisition. That is, it is possible that participants may have learned another language (e.g. Hindi, Bangla, or Marathi in India or Arabic, Spanish, or Chinese in the United States) before K-12. It could also be the case that any of the participants in either country may have learned English at home before K-12.

[8] Because McGurk effects show large individual differences and depend strongly on the stimuli used, an alternate model for the /a_a/ vowel condition was run using a looser criterion where participants who experienced veridical percepts at a rate of chance (33%) in veridical check trials were included. This criterion resulted in, unsurprisingly, a larger number of participants being included and importantly, all effects held in the model tested with looser criteria; see section X.

participant self-reported any history of a speech or hearing disorder diagnosis. Participants were paid $3.25 USD for completing the experimental session.

2.1.2 Stimuli

2.1.2.1 Audio Materials

Audio stimuli were composed of English non-word sequences [ibi] and [aba]. One male speaker of American English and one male speaker of Indian English served as the model talkers for the experiment and recorded the English non-words from a randomized list where they repeated each target item 10 times. Speakers were instructed to produce the non-words with a trochaic stress pattern. From the 10 repetitions, the best production of each non-word, those free of noise or mispronunciations, was selected for inclusion. Audio recordings were made in a sound attenuated booth at the University of Michigan Phonetics Laboratory.

Each model talker was digitally recorded onto a MacBook Pro laptop computer using an AKG C 4000 B microphone and an external Focusrite Scarlet Solo preamplifier. Recordings were made with a sampling rate of 44.1kHz in Praat (Boersma & Weenink, 2022). All tokens selected for inclusion were equalized to have an average intensity of 70dB using the Scale Intensity function in Praat (Boersma & Weenink, 2022). In an effort to neutralize speaker specific artifacts, both socioindexical and otherwise, consonant and vowel durations of included target items were edited to reflect the average vowel and consonant durations across the American English and Indian English model talkers for their [aba] and [ibi] productions[9]. This

[9] Originally, these stimuli were created to be used with electrophysiological dependent measures (e.g. N1/P2). As a result many of the edits controlled for sub-millisecond differences between the speakers.

was carried out by either excising acoustic content from the midpoint of target vowels (where formants were relatively stable) and the closure portion of consonants at the zero crossing or by doubling acoustic content at the midpoint of target vowels and consonant closures at the zero crossing. Table 2.1 provides the durations for each consonant and vowel in the [ibi] and [aba] stimulus recordings as well as the durational differences between the vowels and consonants for each model talker. In both vowel conditions, stimuli showed the clear burst at the release of the [b] constriction (Figure 2.1 for [ibi]).

Stimulus	V1 Duration	C Duration	V2 Duration	Total Duration
Indian English [ibi]	139 ms	61 ms	146 ms	346 ms
American English [ibi]	138 ms	62 ms	147 ms	347 ms
Difference	1 ms	1 ms	1 ms	1 ms
Indian English [aba]	153 ms	62 ms	107 ms	322 ms
American English [aba]	153 ms	63 ms	108 ms	324 ms
Difference	0 ms	1 ms	1 ms	2 ms

Table 2.1: Values for [aba] and [ibi] acoustic stimuli from the American English and Indian English model talkers.

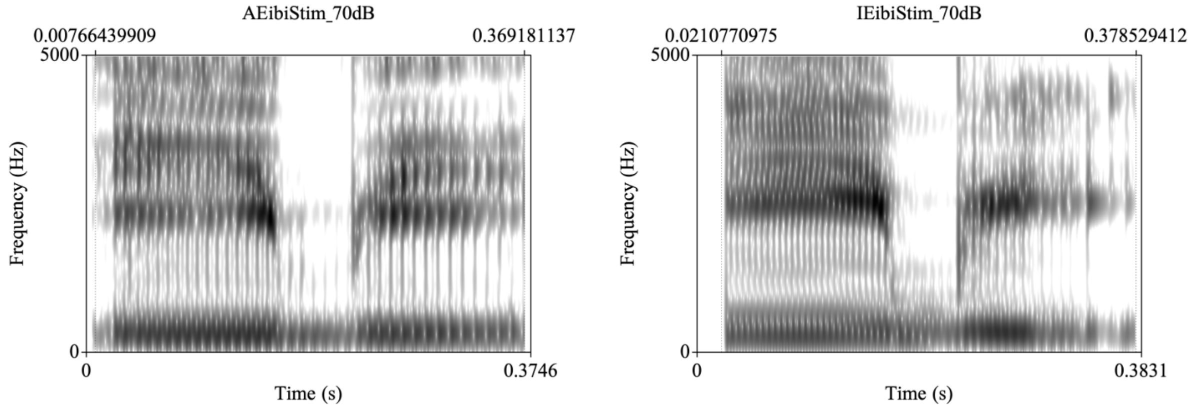

Figure 2.1: American English (left) and Indian English (right) [ibi] stimuli. Note the clear broadband burst of [b] in each spectrogram.

2.1.2.2 Video Materials

Video recordings of each model talker producing target non-word items were digitally recorded onto a Mac Pro using a Canon video camera at a frame rate of 29.97fps (29.97Hz). In each target item the consonant articulation was either [b], [g], [ð], [w], or [v] for the American English model talker and [b], [g], [d̪], [ʋ] for the Indian English model talker[10]. Vowel articulations were either [i] or [a]. In all target items, the first and final vowels were matched ([ibi], [aba], [iwi], [awa], etc.) resulting in 10 combinations (5 consonants x 2 vowels). These recordings were elicited from each model talker via a recording list projected onto a teleprompter in front of the camera. Both model talkers repeated each target item 10 times and from the 10 repetitions the best productions of each non-word, those free from potentially confounding co-gestures (e.g. eye movements away from camera, head tilts) or those with the clearest visual articulations, were selected for inclusion. Figure 2.2 shows a still from the

[10] Both the American English and Indian English model talkers were prompted with the same teleprompter materials asking them to produce orthographically presented syllables like ‘awa’ or ‘idi’. The differences in inventories here between model talkers is a description of the consonants they produced. As noted below in § 2.1.2.3 any acoustic cues that might signal these differences to participants were deleted from the audio channels of the video recordings resulting in similar silent visual articulations. As such, while the Indian English inventory only consists of [ʋ], there were two versions of [ʋ] one prompted by ‘v’ orthography and one prompted by ‘w’ orthography.

included video stimulus /ava/. Video components of the stimuli were made in the Advanced Videocasting Suite at the University of Michigan LSA Media Center.

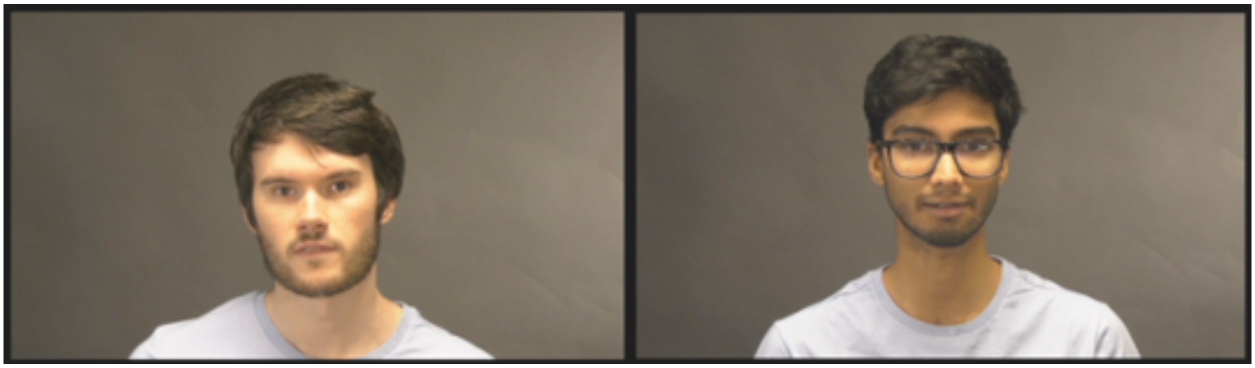

Figure 2.2: Video still of the American English and Indian English model talkers producing [v] (American English) and [ʋ] (Indian English) at maximal closure during an /ava/ target item elicitation. Note the labio-dental closure for the American English model talker and the labio-dental aperture for the Indian English model talker.

2.1.2.3 Audiovisual Materials

After digital audio and video recordings had been obtained from the model talkers, they were synced together using iMovie to make illusory (audiovisual incongruent) and veridical (audiovisual congruent) audiovisual stimuli. The original audio from the video recordings was deleted and the audio stimuli made in the University of Michigan Phonetics Lab (§ 2.1.2.1) were dubbed over the videos. In the case of /b/[11] and /g/, audio and video were synced at the burst release seen in the waveform on the audio channel and the first visual evidence of stop release as seen on the video channel. In the case of /ð/, /v/, and /w/, audio and video were synced at the first instance of intensity drop off as seen in the waveform on the audio channel and the first visual instance of consonantal constriction as seen in the video channel.

[11] Throughout this chapter I will be using bracket notation (e.g. [v]) for the unidimensional audio and visual signals as they have measurable qualities that can be used to describe them. For audiovisual materials, I will be using slash notation (e.g. /v/) as these items are primarily illusory and their perceptual whole as experienced by listeners is not measurable in the same way. This shorthand can be thought of in a similar way that phonetic and phonological descriptions of sounds are often partitioned in the literature. This also means that I will be writing from an American English-centric perspective in the case of illusory stimuli /b/, /d/, /ð/, /v/ and /w/ as these were the orthographic response options offered to participants.

Audiovisual materials then underwent norming with 10 American English speakers to determine which illusory configurations (e.g., visual [v], audio [b]) elicited the strongest illusory (e.g. /v/) response. The configurations that scored the highest for each model talker (American English, Indian English), consonant (/b/, /d/, /ð/, /w/, or /v/; see §1.4, Table 1.3 for American English and Indian English pronunciation differences), and vowel (/a/, /i/) combination were kept, resulting in 20 stimuli (8 illusory + 2 veridical x 2 model talkers). Table 2.2 shows the configurations for building the illusory and veridical audiovisual stimuli.

Audio	AE Visual	IE visual	[i_i]	[a_a]
[b]	[b]	[b]	/ibi/	/aba/
[b]	[g]	[g]	/idi/	/ada/
[b]	[ð]	[d̤]	/iði/	/aða/
[b]	[w]	[ʋ]	/iwi/	/awa/
[b]	[v]	[ʋ]	/ivi/	/ava/

Table 2.2: Audiovisual stimuli created for each of the model talkers. Stimuli designed to elicit illusory percepts are highlighted in the grey cells. All consonants occur in both /i_i/ and /a_a/ contexts.

2.1.2.4 Intervention Video

The intervention video was composed of a six-minute video of the model talkers explaining the differences between transverse and longitudinal waves as well as the physics behind the Doppler effect. The intervention video opens with both model talkers in frame and then moves to individual close ups with animations depicting the effects each talker is describing. In these close-ups, it is always the Indian English model talker followed by the American English model talker before returning to both talkers being in frame. This formulation, Indian English model talker, American English model talker, both talkers together is repeated twice after the opening. Like the visual stimuli, the intervention video was digitally recorded onto a Mac Pro using a Canon video camera at a frame rate of 29.97fps (29.97Hz). The audio for this portion was captured using a single shotgun microphone and digitally recorded using the audio defaults on the Canon video camera. Intervention stimuli were edited in iMovie and wave animations (Russell 2014, Perkins et al. 2006; e.g., Figure 2.3, left and right frame) were added. The intervention stimuli were also made in the Advanced Videocasting Suite at the University of Michigan LSA Media Center. The full script for the intervention video is included in Appendix A.

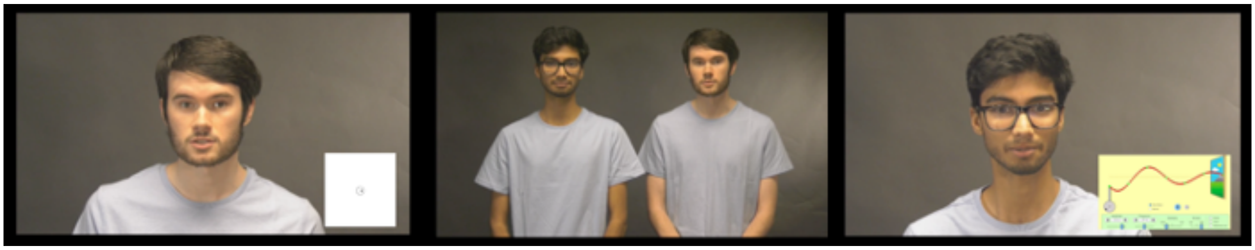

Figure 2.3: Video still from the intervention video. In the leftmost frame, the American English model talker explains the doppler effect with an animation from Russell (2014). In the center frame, the Indian English and American English model talkers welcome the participant to the intervention video. In the rightmost frame, the Indian English model talker explains the mechanics of transverse waves with animation from Perkins et al. (2006).

2.1.3 Procedure

Each participant was tested in a single experimental session from an online location of their choice within India or the United States. Participants accessed the experiment by accepting a HIT (Human Intelligence Task) on the Amazon Mechanical Turk platform and were then directed to the Gorilla Experiment Builder platform (Anwyl-Irvine et al. 2019) where the experiment was hosted. Participants were instructed to wear headphones during the task and completed a consent form and headphone test prior to beginning the experimental session. In the roughly 15-minute session, participants completed a perceptual adaptation task which consisted of two categorization tasks, mediated by the intervention video. Trials were confined to audiovisual conditions to preserve the integrity of the intervention phase. While unimodal trials are often used in illusory perception research to increase the explanatory power of the contributions of the individual acoustic and visual modalities, in this design unimodal trials would have had the disadvantage of giving participants experience with the accents of the model talkers, potentially inducing an adaptive shift in the pre-intervention phase. Because the primary question centers on how experience drives perception, unimodal trial were not used. At the end of all experimental tasks, participants completed a language background questionnaire.

2.1.3.1 Perceptual Adaptation Task

The perceptual adaptation task was divided into 3 phases: 1) pre-intervention categorization, 2) intervention, 3) post-intervention categorization. In both of the categorization phases, the perceptual adaptation task was broken into four blocks, where the vowel (/a_a/ or /i_i/) and consonant (i.e. Labial: [w], [v] or [ʋ] or Coronal[12]: [g], [ð] or [d̤]) type were held constant within a block. These four blocks (e.g. Labial /a_a/, Coronal /a_a/, Labial /i_i/, Coronal /i_i/) were counterbalanced to avoid confounding effects related to test ordering. Both model talkers were included in each block.

In each categorization trial, participants were required to press a play button to begin the stimulus video on their screen. Once started, each video would play (~ 500ms) and then immediately move to the response screen. On this screen, participants chose from 3-alternative-forced choices to “report what consonant [they thought] the person said”. This wording was chosen over, “what you heard” to avoid biasing participants towards one sensory domain over

[12] Coronal is broadly describing the response options that were presented to participants given that they only ever see visual gestural evidence of interdental [ð], dental [d̪ʱ], or velar [g] constrictions in the audiovisual materials.

another. On response screens in the labial blocks, participants could choose from ‘b’, ‘v’, or ‘w’. On the response screens in the coronal blocks, participants could choose from ‘b’, ‘d’, ‘th’. Participants were told that if they weren’t sure what was said to take their best guess. After participants made their choice as to what they thought was said, their response was recorded, the trial ended, and they were moved to the next trial. Categorization responses were collected from both the pre- and post-intervention phases and analyzed.

At the end of the first four categorization blocks, participants were moved to the intervention phase of the experiment. In the intervention phase, participants were instructed to press play on the intervention video and to pay attention to the screen during the intervention phase as they would be asked about the content of the video. Within the doppler effect portion of the intervention video, a video of duck swimming on a pond, demonstrating a visual example of the doppler effect, was included in the animation area. There was a single attention check trial at the end of the intervention phase that asked participants what animal they saw during the intervention. All participants successfully completed this check trial.

2.1.3.2 Questionnaire

At the end of all experimental tasks, participants were asked to complete a language background questionnaire (Appendix B). Participants reported their language background, language usage, and whether they’re perceived by others to have an accent in any of the languages they speak by other speakers of those languages. This last question was included as a secondary metalinguistic way to determine if Indian English speakers who answered that they were speakers of English shared phonological patterns with the Indian English model talker. In addition, participants were asked what they thought the experiment was about. While many participants believed that the experiment was about accent perception and how visual inputs can aid in accent perception, there were no reports about the stimuli being ‘odd’ or ‘unnatural’ nor was the McGurk effect named.

2.1.3.3 Statistical Analysis

As laid out above, the two research questions being investigated in this experiment are: 1) Does the proportion of illusions experienced by participants change after gaining experience with the accent of the model talker? and 2) Do categorization rates of illusory percepts, about the type of percept a participant experienced, change after participants gain experience with the accent of the model talkers? To investigate these questions, Bayesian Binomial Logistic Regressions were run using the `brms` package in R to model participant responses. Statistical models were run for each of the four blocks (i.e. Labial /a_a/, Coronal /a_a/, Labial /i_i/, Coronal /i_i/) given that participants were only ever presented a subset of possible responses depending on the block (i.e. /b, v, w/ responses in Labial blocks and /b, d, ð/ responses in Coronal blocks) and given that individual participants experienced illusory percepts at different rates across the blocks (e.g. any given participant may have experienced illusory percepts at a rate greater than chance in labial conditions but not coronal conditions). For models targeting the proportion of illusions experienced by participants (i.e. illusory effectiveness models), the dependent variable was whether the reported percept reflected the auditory signal or not (1-illusion, 0-veridical). The effects of intervention (pre/post), model talker (American English, Indian English), listener group (American English, Indian English), and visual articulation construed broadly (/ð/ comprised of [ð], [ḍ̤], /d/ comprised of [g], /v/ comprised of [v], [ʋ] or /w/ comprised of [w], [ʋ]), their interactions, and a random intercept for participant were included in the model.

For models targeting the categorization of illusory percepts experienced by participants (i.e. illusory categorization models), the dependent variable was whether the reported illusory percept was a fricative (e.g. /v/ or /ð/) or not (e.g. /w/ or /d/). Like illusory effectiveness models, the effects of intervention (pre/post), model talker (American English, Indian English), listener group (American English, Indian English), and visual articulation construed broadly (/ð/ comprised of [ð], [ḍ̤], /d/ comprised of [g], /v/ comprised of [v], [ʋ], and /w/ comprised of [w] and [ʋ]), their interactions, and a random intercept for participant were included in the model. Illusory categorization models only include trials in which a participant's response was consistent with experiencing an illusion as opposed to illusory effectiveness models where all experimental trials are included.

2.2 Predictions

2.2.1 American English Pre-Intervention

American English participants are expected to come to the task with separate phonemic categories for /d/, /ð/, /v/, and /w/. In the case of /d/, /ð/, and /v/ there are very clear visual articulations within American English that facilitate illusory assignment of the burst on [b] to one of these three categories. In the case of [v] this is the articulation of the upper teeth touching the bottom lip, in [ð] the tongue protruding between the teeth, and in [d] movement of the tongue behind the teeth in conjunction with a lack of bilabial closure – all of which have been shown to facilitate illusory perception in American English listeners (McGurk & MacDonald 1976; MacDonald & McGurk 1978; Green, Kuhl, & Metzloff 1988; Green &

Gerdeman, 1995; Green, Kuhl, Meltzoff, & Stevens 1991; Rosenblum & Saldaña 1996; Green & Norrix 1997; Brancazio et al. 2003; Wang et al. 2008; Tiipaana 2014; Rosenblum 2019).

Unlike the three cases presented above, there is no frication involved in the articulation of /w/ in American English and, thus, the burst of [b] has nothing to map onto when paired with a visual [w] articulation. This creates a division that I exploit when probing how American English listeners perceive illusory stimuli from their own dialect. Thus, as outlined below in Table 2.3, I expect that listeners will perceive /v/ when viewing [v], /d/ when viewing [g], and /ð/ when viewing [ð] while shadowing the American English model talker. I also expect that American English listeners will fail to experience an illusory percept when viewing [w] and accordingly will perceive [b] while listening to the American English model talker.

While the expectations of American English participants should match those of the American English model talker, given their shared phonologies, Indian English model talker productions will require American English participants to confront articulations that do not inherently map as transparently to the /v/, /w/, /d/, and /ð/ categories of American English. This is particularly relevant in the case of [ʋ], which exhibits the lip rounding associated with /w/ in American English as well as the frication generated from a constriction between the upper teeth and lower lip associated with /v/. Likewise, Indian English lacks the interdental fricative /ð/ found in American English but does maintain a distinction between dental and retroflexed stops. The Indian English model talker in this study reliably generated a voiced aspirated dental stop [ḍ̤] with clear visual articulations where the tongue tip meets the bottom of the front teeth when he was asked to produce [iði] or [aða] during stimulus creation. For both the [ʋ] and [ḍ̤] articulations, there is sufficient visual articulatory evidence, particularly when paired with the audio for [b] that American English participants should perceive the fricatives /v/ and /ð/.

Given that American English participants will be unexposed to the actual phonological patterns of the Indian English model talker at this stage of the task, and as such will not need to perceptually adapt their dialect pre-intervention, I expect that American English listeners will perceive /d/ when viewing [g], /ð/ when viewing the tongue to tooth contact of [d̤], and /v/ when viewing the upper tooth to lower lip constriction of [ʋ]. Table 2.3 shows the expected percepts for American English speaking participants given different audio visual configurations.

Audio	Visual (AE Talker)	Visual (IE Talker)	Predicted Percept (AE)	Predicted Percept (IE)
[b]	[v]	[ʋ]	/v/	/v/
[b]	[w]	[ʋ]	/b/	/v/
[b]	[g]	[g]	/d/	/d/
[b]	[ð]	[d̪d̤]	/ð/	/ð/

Table 2.3: Predicted percepts given audiovisual stimulus configurations for American English participants pre-intervention.

2.2.2 Indian English Pre-Intervention

Indian English participants are expected to come to the task with separate phonemic categories for /d/ and /ʋ/. Unlike American English, for Indian English participants there is potentially a wider range of visual articulations that could be associated with an illusion's membership in the /d/ and /ʋ/ phonological categories given the visual stimuli in this experiment. In the case of the [d̤] and [g] articulations produced by the Indian English model talker, we can expect to see a predominance of /d/ responses from Indian English participants given that, considering the absence of a /ð/ phoneme in Indian English, all visual articulations for [g] and [d̤], when paired with acoustic [b], should map back to /d/. In the case of [ʋ], visual articulations should facilitate the mapping the burst of [b] onto the /ʋ/ phonological category.

However, given the response options available to participants in labial trials (e.g. ‘b’, ‘v’, and ‘w’) and because Indian English merges American English /v/ and /w/ into the single category /ʋ/, we can expect to see categorization patterns that look like performance at chance as detailed in Table 2.4 below. Like American English participants pre-intervention, Indian English participants are predicted to use Indian English phonological categories as perceptual targets for all pre-intervention trials given the lack of experience with the accents of the model talkers.

Audio	Visual (AE Talker)	Visual (IE Talker)	Predicted Percept (AE)	Predicted Percept (IE)
[b]	[v]	[ʋ]	/v/ or /w/	/v/ or /w/
[b]	[w]	[ʋ]	/v/ or /w/	/v/ or /w/
[b]	[g]	[g]	/d/	/d/
[b]	[ð]	[ḍ̤]	/ð/	/d/

Table 2.4: Predicted percepts given audiovisual stimulus configurations for Indian English participants pre-intervention.

Akin to how American English participants confront articulations that do not transparently map to their phonological categories when observing the Indian English model talker, Indian English participants also must confront how to map the [g], [ð], [v], and [w] articulations of the American English model talker onto their own /ʋ/ and /d/ categories. However, in pre-intervention this may be somewhat more straight forward for Indian English participants given that a wider range of visual articulations can be mapped to fewer number of phonological categories. Like with the Indian English model talker, [g] productions by the American English model talker should straightforwardly be perceived as /d/ by Indian English participants. Additionally, given the lack of a phonemic category for /ð/ in Indian English, productions of [ð] should also be categorized as /d/ given their articulatory similarity to consonants like [d̤] in Indian English. With categorizations of [v] and [w], again, because of their merger into /ʋ/ in Indian English I expect that Indian English participants will categorize [v] and [w] productions by the American English model talker with patterns that resemble guessing at chance.

It is worth noting that there is the possibility that the cases of [ð], [v], and [w] could present a more complicated categorization pattern than the picture painted above. Despite the experimental stimuli being non-words and the participants having no experience with the actual dialectal productions of the model speakers in the pre-intervention phase, Indian English is often seen as a 'marked' variety in the array of global Englishes. As such, speakers of this English are often stigmatized and are made aware of their 'non-standard' productions particularly in inter-dialectal settings where speakers of the same language can rely on shibboleths and the semantic content of real words to aid in perceptually adapting to one another. Often the linguistic ideologies that come with these corrections of 'non-standard' patterns lift up the phonological patterns spoken by homeland or colonizer populations (e.g. British English, American English, Australian English, New Zealand English) as the 'ideal' target pronunciations while at the same time subjugating the phonological patterns spoken by colonized and marginalized homeland/immigrant populations (e.g. Indian English, African American Vernacular English, various Caribbean and African Englishes). As noted in § 1.4 there is a raciolinguistic dimension to the sociolinguistic reality that Indian English speakers face and Indian English participants could utilize their experience in that reality in this task. While it is likely the case that American English participants have experience interacting with racially Indian presenting individuals who use either American English or Indian English pronunciations, it is much less likely that Indian English participants have experience with racially white presenting American English individuals using Indian English pronunciations. Were Indian English participants to lean on this experience, I would expect that Indian English participants would align more closely with American English participants when listening to the American English model talker in pre-intervention, as laid out in Table 2.4, and as a result show a weaker adaptation pattern in post-intervention.

2.2.3 Post-Intervention Predictions

After viewing the intervention video, participants of both dialect groups (Indian English and American English) will have experience with the accents of the model talkers. In line with the perceptual adaptation literature (Bertelson et al. 2003, Norris, McQueen & Cutler 2003, Kraljic & Samuel 2006, Trude & Brown-Schmidt 2012, Reinisch & Mitterer 2016), participants l update their perceptual strategies to include meaningful socio-indexical information about the speech patterns of both of the model speakers. In the case of across-dialect talker information, I expect the illusory categorization rates to change as a function of this phonological updating. In the case of within-dialect perceptual updating, this shouldn't change the illusory categorization rates, given the specific experimental stimuli. While participants may adapt to speaker-specific within-category pronunciations, this shouldn't change which stimuli are considered members of a category. Broadly, for a given participant group, pre-to-post intervention changes should be in the direction of the other groups' pre-intervention pattern. Table 2.5 lays out the predicted changes in perception for American English and Indian English participants.

Model Talker	Participant	Audio	Visual	Change
American English	Indian English	[b]	[v]	Increase in /v/
American English	Indian English	[b]	[w]	Increase in /b/ (Failure to McGurk)
American English	Indian English	[b]	[g]	No Change
American English	Indian English	[b]	[ð]	Increase in /ð/
Indian English	American English	[b]	[ʋ]	Decrease in /v/
Indian English	American English	[b]	[ʋ]	Decrease in /w/
Indian English	American English	[b]	[g]	No Change
Indian English	American English	[b]	[d̤]	Decrease in /ð/

Table 2.5: Changes in predicted percepts given audiovisual stimulus configurations after intervention. Indian English participants are expected to shift toward American English performance in pre-intervention and American English participants are expected to shift toward pre-intervention Indian English performance.

For American English participants I predict no pre-to-post intervention change in categorization rates when listening to the American English model talker. However, when listening to the Indian English model talker I expect changes in /v/, /w/, and /ð/ categorization. Because of the near-merger of [v] and [w] to [ʋ] in Indian English from an American English perspective, I would expect categorization rates of /v/ and /w/ to approach chance if American English participants update their perceptual strategies to reflect their experience in the intervention phase. I also expect to see a decrease in /ð/ categorization rates for American English participants given what would appear to be a pattern of substitution /ð/ → [d] in the intervention video for the Indian English model talker. In the case of [g] being categorized as /d/ I expect to see no change.

For Indian English participants I predict no change in categorization rates when listening to the Indian English model talker. However, when listening to the American English model talker I expect changes in /v/, /w/, and /ð/ categorization. Because of the phonemic distinction between /v/ and /w/ in American English, Indian English participants must split their /ʋ/ category if they're to update their perceptual strategy to reflect their experience in the intervention phase. While "merged" listeners often do not hear distinctions in other dialects for their merged variants, the /v, w/ merger is highly salient and meta-linguistically marked for Indian English listeners (Sailaja 2012). In a similar vein, I would expect to see an increase in /ð/ categorization rates for Indian English participants, again given the distinction for the American English model talker. However, given that /ð/-stopping is more widely attested in global Englishes in addition to not being as marked as the /v, w/ merger in Indian English, the increase in /ð/ responses might not be as extreme. As was the case with American English participants I expect to see no change in /d/ categorization when Indian English participants are presented with a visual articulation of [g].

In addition to changes in categorization rates for the identity of illusory stimuli, I also expect illusion rates to increase for almost all visual articulations after intervention for participants in both language groups. The exception to this is [w] with American English participants, which should result in an illusion rate decrease.

2.3 Results

2.3.1 Results for /a_a/ context, coronal consonants

Within the Coronal /a_a/ block, there were 133 (out of a total of 200) participants (69 American English, 64 Indian English) who experienced illusory percepts 33% or more of the time (i.e., more often than chance). These participants' results were included in the analyses. In all conditions, participants experienced more illusory percepts than veridical percepts (Figure 2.4, the left vs. right paired columns of each panel). The illusory effectiveness model revealed a main effect[13] of visual articulation (β = -0.92, CI = -1.51, -0.36) where participants were more likely to experience and illusion when the veridical articulation was [ð]/[d̤] as opposed to [g]. While there was a numerical trend for increased illusions after the intervention, this main effect was not statistically reliable. Pairwise comparisons were conducted using the `emmeans` package in R to investigate whether there was an effect of intervention in a subset of conditions. All pairwise comparisons were adjusted with Tukey corrections for multiple comparisons. Those comparisons show that participants experienced more illusions after intervention in five of the eight conditions (Table 2.6; starred rows).

Figure 2.4 visualizes these differences by showing the number of illusory responses according to participant group (rows) and model talker and visual articulation (columns). Both American English and Indian English participants showed an effect of intervention when the visual articulation was [g], as seen by the increase in illusory percepts post-intervention (Figures 2.4c, 2.4d, 2.4g, 2.4h). Additionally, pairwise comparisons show that Indian English

[13] I interpret as 'statistically reliable' effects where the β-value is not equal to zero and whose 95% credibility interval excludes zero.

participants also had more illusions after intervention when listening to an Indian English model talker with a [d̤] visual articulation (Figure 2.4f).

Intervention	Visual Articulation	Model Talker	Group	Est/Lower CI/Upper CI
pre/post	[ð]	AE	AE	-0.28428 -0.87570 0.3441
pre/post	[g]	AE	AE	-0.81345 -1.40033 -0.2991 *
pre/post	[d̤]	IE	AE	-0.49685 -1.13327 0.1771
pre/post	[g]	IE	AE	-1.23000 -1.81096 -0.6485 *
pre/post	[ð]	AE	IE	-0.50806 -1.13980 0.1531
pre/post	[g]	AE	IE	-0.83536 -1.39773 -0.2573 *
pre/post	[d̤]	IE	IE	-0.86693 -1.56791 -0.2015 *
pre/post	[g]	IE	IE	-0.69273 -1.34722 -0.1129 *

Table 2.6: Pairwise comparisons for effects of intervention on illusory effectiveness in coronal /a_a/ blocks. Comparisons for which there is an effect are starred.

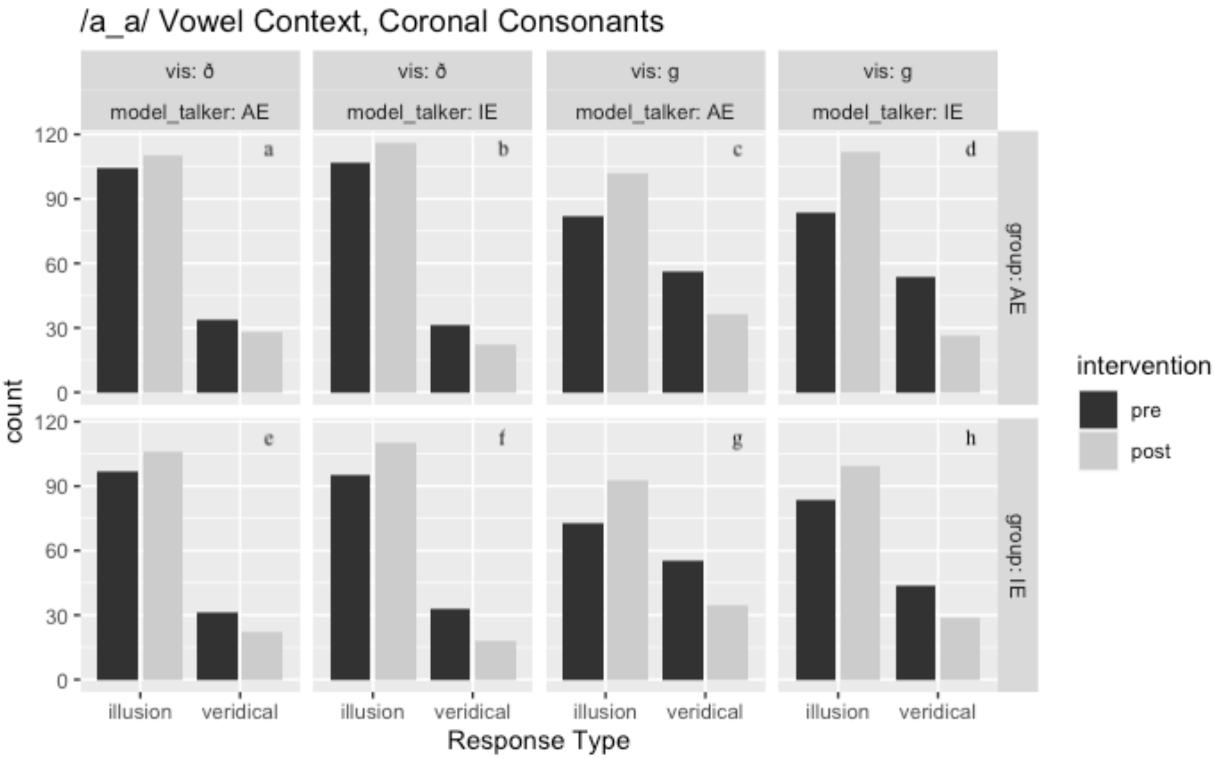

Figure 2.4: Rates of illusory effectiveness in the coronal /a_a/ block. Each panel of the figure shows the counts of illusory and veridical responses for a particular visual articulation (vis: ð, g), model talker (talker: AE, IE), and participant group (group: AE, IE) condition.

Figure 2.5 visualizes the results of the categorization model for the Coronal /a_a/ block. The illusory categorization model revealed main effects of model talker (β = -0.92 , CI = -1.58, -0.26) and visual articulation (β = -2.33, CI = -3.12, -1.60). These two effects interacted (β = 1.46, CI = 0.48, 2.49) such that participants were more likely to report perceiving /d/ (green bars) when viewing the Indian English model talker producing [d̤] than the American English talker producing [ð] (compare Figure 2.5a and 2.5b, and Figure 2.5e and 2.5f).

Tukey-adjusted pairwise comparisons were conducted to investigate whether there was an effect of intervention, but, against expectations, pairwise comparisons revealed no effects of intervention, as seen in Table 2.7 and Figure 2.5 (pre and post comparisons).

Intervention	**Visual Articulation**	**Model Talker**	**Group**	**Est/Lower CI/Upper CI**
pre/post	[ð]	AE	AE	-0.48909 -1.1847 0.2042
pre/post	[g]	AE	AE	-0.57815 -1.3254 0.1234
pre/post	[d̤]	IE	AE	-0.41831 -1.0177 0.2178
pre/post	[g]	IE	AE	0.32549 -0.3348 0.9972
pre/post	[ð]	AE	IE	0.24837 -0.3965 0.8772
pre/post	[g]	AE	IE	-0.36656 -1.1122 0.3872
pre/post	[d̤]	IE	IE	0.08321 -0.5091 0.6814
pre/post	[g]	IE	IE	-0.34554 -1.0509 0.3368

Table 2.7: Pairwise comparisons for effects of intervention on illusory categorization in coronal /a_a/ blocks.

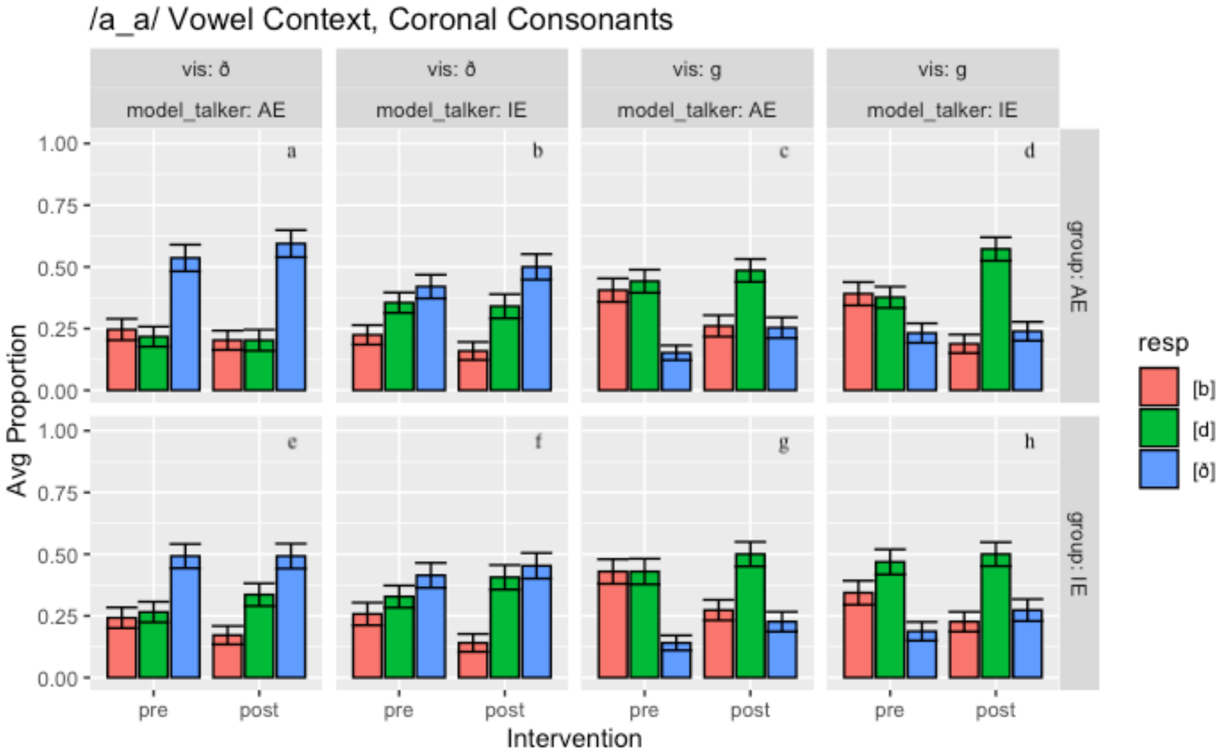

Figure 2.5: Rates of illusory categorization in the coronal /a_a/ block. Each panel of the figure shows the proportion of categorized illusory responses for a particular visual articulation (vis: ð, g), model talker (speaker: AE, IE), participant group (group: AE, IE) and the standard error.

2.3.2 Results for /i_i/ context, coronal consonants

Within the Coronal /i_i/ block, 134 participants (67 American English, 67 Indian English) experienced illusory percepts 33% or more of the time and were included for analysis. In all conditions, there were more illusory percepts than veridical percepts (Figure 2.6). The illusory effectiveness model revealed a main effect of visual articulation (β = -0.65, CI = -1.24, -0.08), mediated by an interaction of model talker and visual articulation (β = 1.16, CI = 0.33, 2.04) such that participants were more likely to experience illusions when the Indian English model talker produced a visual articulation of [g] than when the American English model talker did (Figure 2.6d, 2.6h vs. Figure 2.6c, 2.6g). While there was no overall main effect of intervention, the results of Tukey-adjusted pairwise comparisons showed that the expected effect of increased illusions after intervention held when the visual articulation was [g] and the talker was the American English model talker (Table 2.8; starred rows). The intervention effect in these conditions is visualized in Figure 2.6, where the illusory responses of both American English participants (Figure 2.6c) and Indian English participants (Figure 2.6g) increase post-intervention, and veridical responses correspondingly decrease.

Intervention	Visual Articulation	Model Talker	Group	Est/Lower CI/Upper CI
pre/post	[ð]	AE	AE	-0.63145 -1.2871 0.00695
pre/post	[g]	AE	AE	-1.05620 -1.6396 -0.43521 *
pre/post	[d̤]	IE	AE	0.03800 -0.5822 0.57314
pre/post	[g]	IE	AE	0.23185 -0.3544 0.85369
pre/post	[ð]	AE	IE	-0.27571 -0.8548 0.33769
pre/post	[g]	AE	IE	-0.79769 -1.3723 -0.22126 *
pre/post	[d̤]	IE	IE	-0.48627 -1.1269 0.09251
pre/post	[g]	IE	IE	-0.32548 -0.9405 0.31282

Table 2.8: Pairwise comparisons for effects of intervention on illusory effectiveness in coronal /i_i/ blocks. Comparisons for which there is an effect are starred.

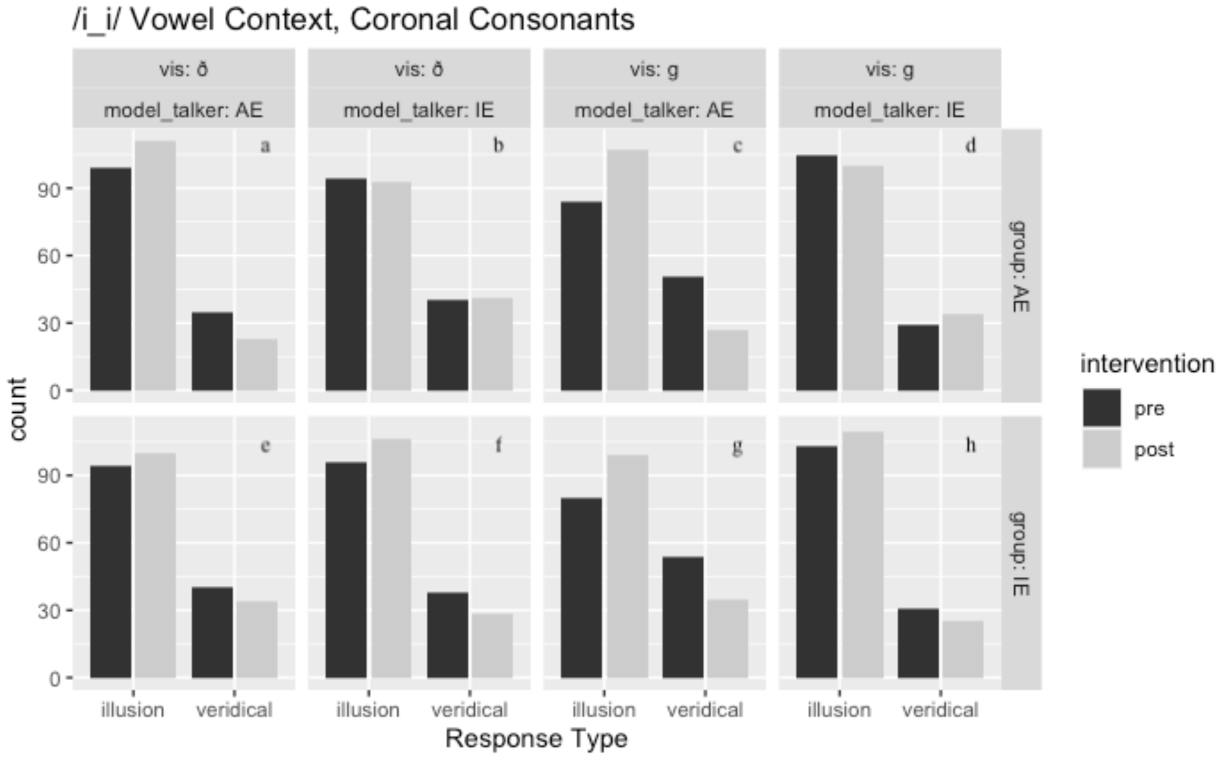

Figure 2.6: Rates of illusory effectiveness in the coronal /i_i/ block. Each panel of the figure shows the counts of illusory and veridical responses for a particular visual articulation (vis: ð, g), model talker (speaker: AE, IE), and participant group (group: AE, IE) condition.

Figure 2.7 visualizes the results of the categorization model for the Coronal /i_i/ block. The illusory categorization model revealed a main effect of intervention (β = 0.78, CI = 0.12, 1.40) such that participants were more likely to experience illusory percepts after the intervention phase. There was also a main effect of visual articulation (β = -0.79, CI -1.46, -0.14) such that participants were more likely to categorize illusory percepts as /ð/ when the visual articulation was [ð]/ [d̤] (Figure 2.7a, 2.7b, 2.7e, 2.7f) as compared to [g] (Figure 2.7c, 2.7d, 2.7g). Tukey-adjusted pairwise comparisons showed that the expected effect of intervention held within groups when the visual articulation was [ð]/[d̤] (Table 2.9; starred rows). For American English participants, their rate of /ð/ responses increased when responding to the American English model talker after intervention (Figure 2.7a). For Indian English participants, their rate of /ð/ responses increased when responding to the Indian English model talker after intervention (Figure 2.7f).

Intervention	**Visual Articulation**	**Model Talker**	**Group**	**Est/Lower CI /Upper CI**
pre/post	[ð]	AE	AE	-0.7808 -1.41340 -0.17750 *
pre/post	[g]	AE	AE	-0.2076 -0.89911 0.44509
pre/post	[ḍ̤]	IE	AE	-0.0384 -0.70565 0.56885
pre/post	[g]	IE	AE	0.3818 -0.33632 1.06807
pre/post	[ð]	AE	IE	-0.1087 -0.73049 0.52778
pre/post	[g]	AE	IE	0.0594 -0.60281 0.73630
pre/post	[ḍ̤]	IE	IE	0.9151 0.32251 1.51868 *
pre/post	[g]	IE	IE	-0.4922 -1.11013 0.08267

Table 2.9: Pairwise comparisons for effects of intervention on illusory categorization in coronal /i_i/ blocks. Comparisons for which there is an effect are starred.

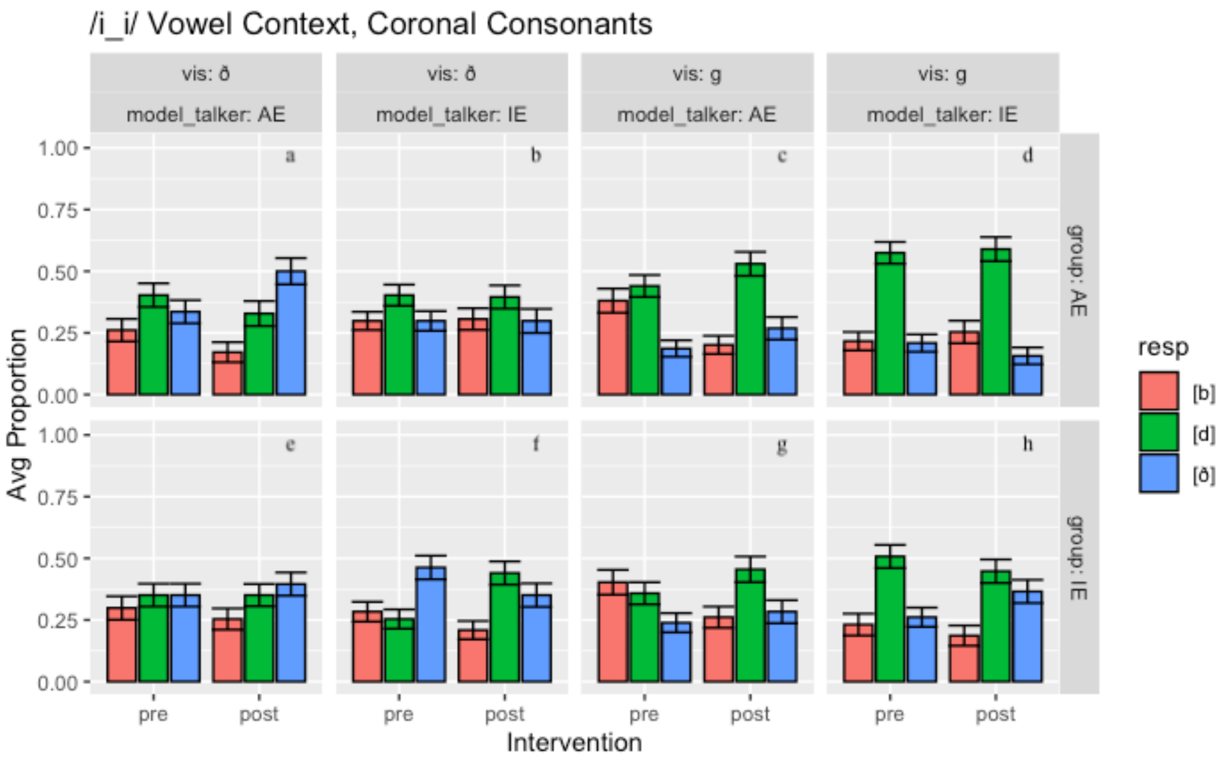

Figure 2.7: Rates of illusory categorization in the coronal /i_i/ block. Each panel of the figure shows the proportion of illusory and veridical responses for a particular visual articulation (vis: ð, g), model talker (speaker: AE, IE), participant group (group: AE, IE) condition and the standard error.

2.3.3 Results for /a_a/ context, labial consonants

Within the Labial /a_a/ block, there were 164 participants (81 American English, 83 Indian English) who experienced illusory percepts 33% or more of the time and whose results were included for analysis. In all conditions, there were more illusory percepts than veridical percepts (Figure 2.8). The illusory effectiveness model revealed a main effect of visual articulation (β = -1.73, CI = -2.28, -1.19) such that participants were more likely to experience an illusion when the visual articulation was [v] (Figure 2.8a, 2.8b, 2.8e, 2.8f) as opposed to [w] (Figure 2.8c, 2.8d, 2.8g, 2.8h). Tukey adjusted pairwise comparisons revealed that there were more illusions after intervention in four conditions (Table 2.10; starred rows). Figure 2.8 visualizes these differences. For American English participants, illusions increased post-intervention when the visual articulation was [w] regardless of model talker (Figure 2.8c and 2.8d), and when the visual articulation was [v] with the Indian English model talker (Figure 2.8b). For Indian English participants, illusions increased post-intervention when the visual articulation was [w] with the Indian English model talker (Figure 2.8h).[14]

Intervention	Visual Articulation	Model Talker	Group	Est/Lower CI/Upper CI
pre/post	[v]	AE	AE	-0.62793 -1.2987 0.03401
pre/post	[w]	AE	AE	-0.67565 -1.1835 -0.20755 *
pre/post	[ʋ] ‘v’	IE	AE	-0.76026 -1.3408 -0.13960 *
pre/post	[ʋ] ‘w’	IE	AE	-0.67314 -1.1814 -0.21165 *
pre/post	[v]	AE	IE	0.11642 -0.5350 0.75536
pre/post	[w]	AE	IE	0.15391 -0.2993 0.65020
pre/post	[ʋ] ‘v’	IE	IE	-0.24912 -0.7632 0.34462
pre/post	[ʋ] ‘w’	IE	IE	-0.60779 -1.0571 -0.09902 *

Table 2.10: Pairwise comparisons for effects of intervention on illusory effectiveness in labial /a_a/ blocks. Comparisons for which there is an effect are starred.

[14] This model was also run using the relaxed criteria described on page 29. Results of the model with relaxed criteria revealed a similar effect of visual articulation ($\beta = -2.35$ CI = -2.92, -1.8) as described above. Tukey adjusted pairwise comparisons were also run for the relaxed criteria model and effects were found in each of the starred conditions in Table 2.10.

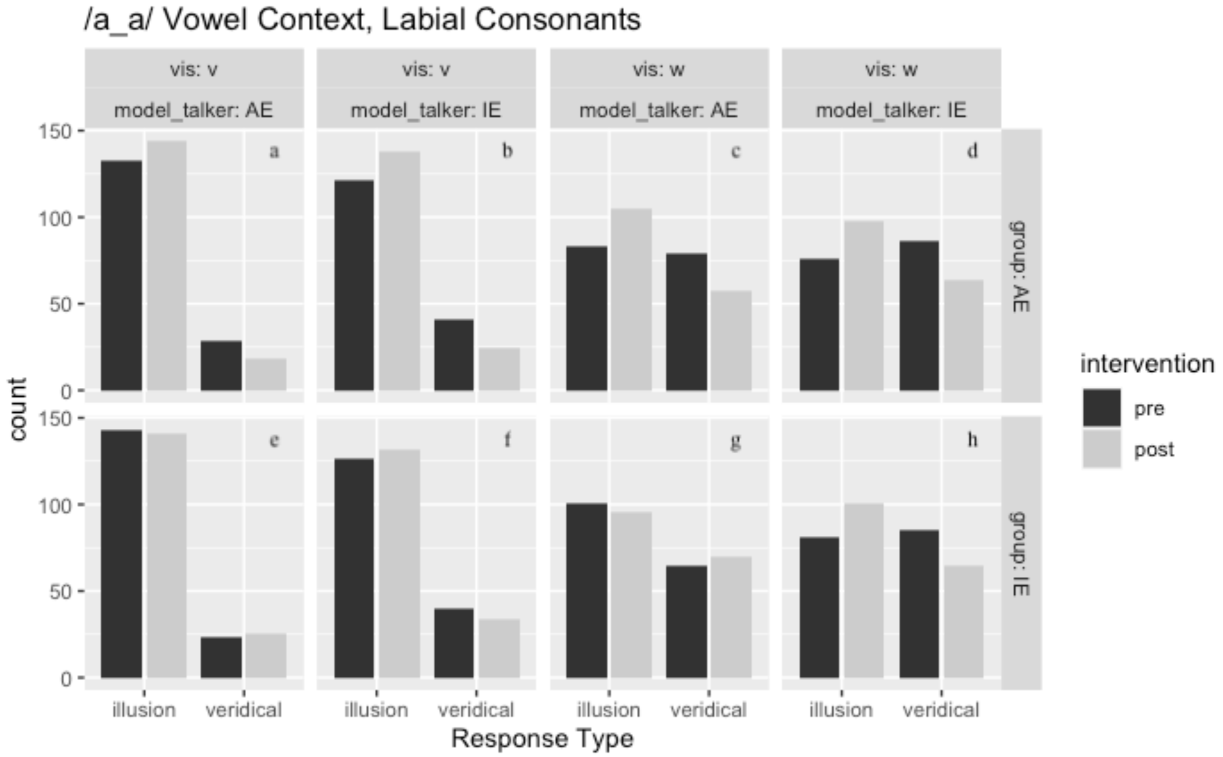

Figure 2.8: Rates of illusory effectiveness in the labial /a_a/ block. Each panel of the figure shows the counts of illusory and veridical responses for a particular visual articulation (vis: v, w), model talker (speaker: AE, IE), and participant group (group: AE, IE) condition.

Figure 2.9 visualizes the results of the categorization model for the Labial /a_a/ block. The illusory categorization model revealed a main effect of visual articulation (β = -1.06, CI = -1.77, -0.38) such that participants were more likely to categorize illusory percepts as /v/ when the visual articulation was [v] (Figure 2.9a, 2.9b, 2.9e, 2.9f) as compared to [w] (Figure 2.9c, 2.9d, 2.9g, 2.9h). There was also a three-way interaction of intervention, model talker, and visual articulation (β = 1.53, CI = 0.18, 2.91) such that both participant groups were more likely to categorize the Indian English model talker's stimuli with visual articulation [ʋ]/'w' as /v/ after intervention (Figure 2.9d and Figure 2.9h). Paired with the results from the illusory effectiveness model, this suggests that both American English and Indian English participants are not only having more illusions with the Indian English talker when he produces [ʋ]/ 'w' visual articulations, as predicted, but participants are categorizing [ʋ]/'w' visual articulations as /v/ suggesting that they're sensitive to the merger found in the actual [ʋ] productions of the Indian English model talker and shifting their response patterns in turn.

While the predicted overall main effect of intervention again was not found, Tukey-adjusted pairwise comparisons did reveal more illusions after intervention for two conditions, both involving American English participants viewing the American English model talker—that is, both involving within-dialect perceptual updating (Table 2.11; starred rows). This is seen in Figure 2.9a and Figure 2.9c. In the case of visual articulation [v], American English participants unexpectedly showed an increase in /v/ categorizations after intervention (Figure 2.9a). In the case of visual articulation [w], American English participants unexpectedly showed an increase in /w/ categorizations (Figure 2.9c). This result will be returned to in §2.4.2 for discussion. [15]

[15] The categorization model was also run using the relaxed criteria described on page 29. Results of the categorization model with relaxed criteria revealed a similar main effect of visual articulation (β = -1.10 CI = -1.86, -0.36) as described above. It also revealed similar three-way interaction between intervention, model talker, and visual articulation as described above (β = 1.74, CI = 0.33, 3.23). Tukey adjusted pairwise comparisons were also run for the relaxed criteria model and effects were found in each of the starred conditions in Table 2.11.

Intervention	Visual Articulation	Talker	Group	Est/Lower CI/Upper CI
pre/post	[v]	AE	AE	-0.6734 -1.42778 -0.00203 *
pre/post	[w]	AE	AE	0.7366 0.11877 1.41851 *
pre/post	[ʋ] ‘v’	IE	AE	0.0387 -0.64464 0.70379
pre/post	[ʋ] ‘w’	IE	AE	-0.0734 -0.74860 0.58874
pre/post	[v]	AE	IE	-0.2163 -0.78501 0.35382
pre/post	[w]	AE	IE	0.1728 -0.39588 0.80251
pre/post	[ʋ] ‘v’	IE	IE	-0.0638 -0.66644 0.51435
pre/post	[ʋ] ‘w’	IE	IE	-0.2343 -0.87707 0.43682

Table 2.11: Pairwise comparisons for effects of intervention on illusory categorization in labial /a_a/ blocks. Comparisons for which there is an effect are starred.

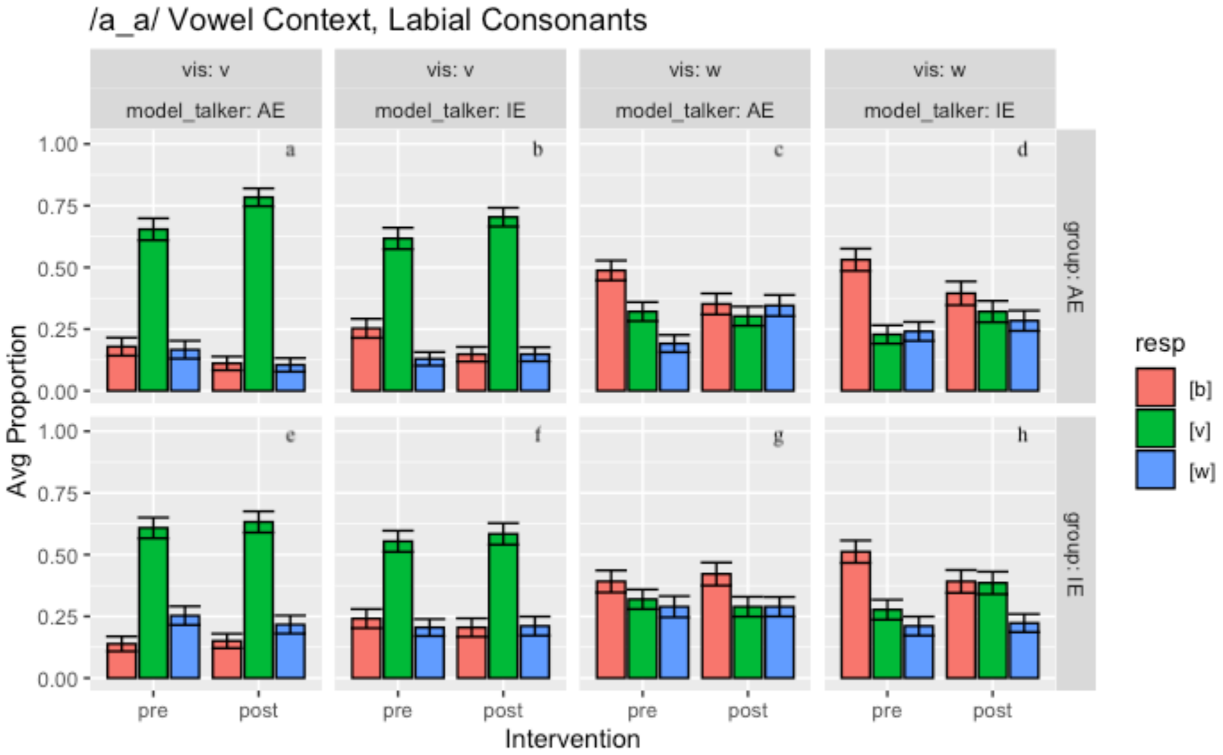

Figure 2.9: Rates of illusory categorization in the labial /a_a/ block. Each panel of the figure shows the proportion of illusory and veridical responses for a particular visual articulation (vis: v, w), model talker (speaker: AE, IE), participant group (group: AE, IE) condition and the standard error.

2.3.4 Results for /i_i/ context, labial consonants

Within the ‘Labial /i_i/’ block, 167 participants (83 American English, 84 Indian English) experienced illusory percepts 33% or more of the time and were included for analysis. In all conditions, there were again more illusory percepts than veridical percepts (Figure 2.10). The illusory effectiveness model revealed a main effect of visual articulation (β = -1.74, CI = -2.33, -1.19) such that participants were more likely to experience an illusion when the visual articulation was a [v] (Figure 2.10a, 2.10b, 2.10e, 2.10f) as opposed to [w] (Figure 2.10c, 2.10d, 2.10g, 2.10h). There was no overall main effect of intervention and Tukey-adjusted pairwise comparisons also revealed no effects of intervention in any of the subset conditions (Table 2.12).

Intervention	**Visual Articulation**	**Model Talker**	**Group**	**Est/Lower CI/Upper CI**
pre/post	[v]	AE	AE	-0.03085 -0.7384 0.595
pre/post	[w]	AE	AE	-0.25471 -0.7275 0.257
pre/post	[ʋ] ‘v’	IE	AE	0.05266 -0.4989 0.621
pre/post	[ʋ] ‘w’	IE	AE	-0.21866 -0.7042 0.278
pre/post	[v]	AE	IE	-0.00549 -0.5716 0.657
pre/post	[w]	AE	IE	-0.08053 -0.5511 0.396
pre/post	[ʋ] ‘v’	IE	IE	0.11193 -0.5189 0.775
pre/post	[ʋ] ‘w’	IE	IE	-0.18054 -0.6845 0.296

Table 2.12: Pairwise comparisons for effects of intervention on illusory effectiveness in labial /i_i/ blocks.

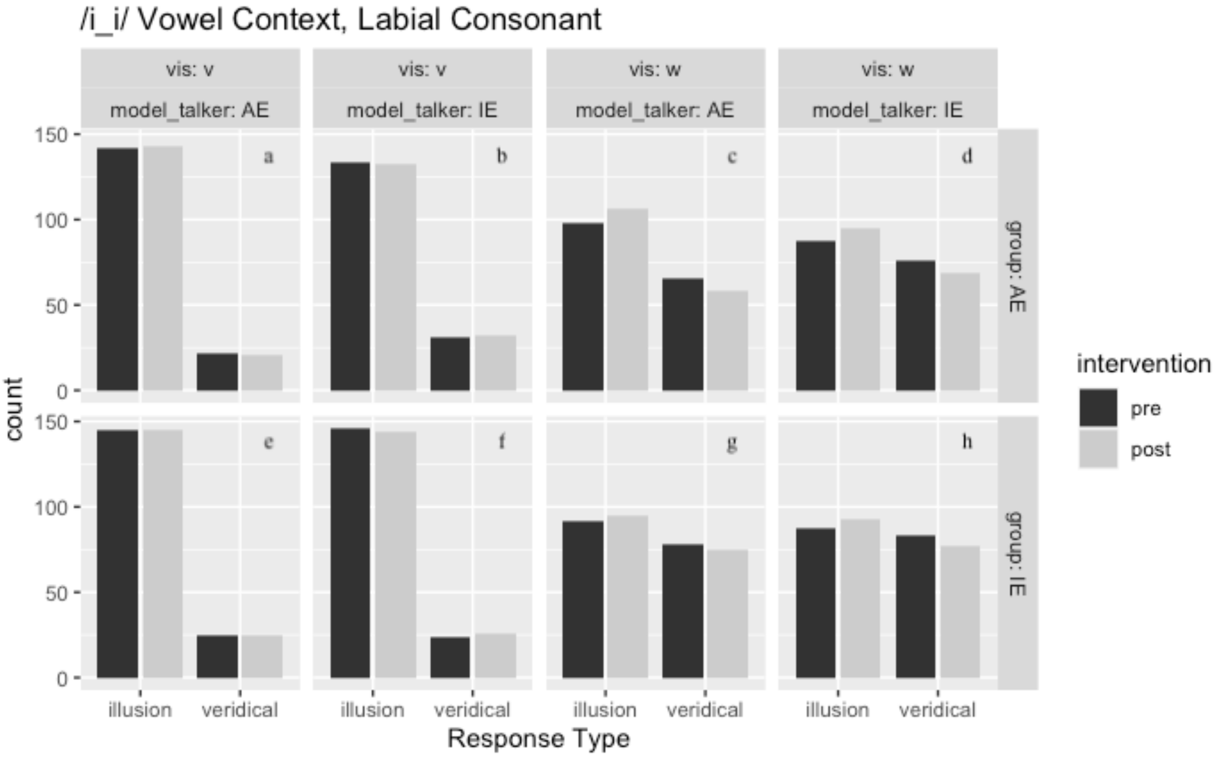

Figure 2.10: Rates of illusory effectiveness in the labial /i_i/ block. Each panel of the figure shows the counts of illusory and veridical responses for a particular visual articulation (vis: v, w), model talker (speaker: AE, IE), and participant group (group: AE, IE) condition.

Figure 2.11 visualizes the proportion of illusory responses for the /i_i/ labial categorization model. The model revealed a main effect of visual articulation (β = -1.67, CI = -2.32, -1.02) such that participants were more likely to categorize illusory percepts as /v/ when the visual articulation was [v]/[ʋ] (2.11a, 2.11b, 2.11e, 2.11f) than when it was [w]/[ʋ] (Figure 2.11c, 2.11d, 2.11g, 2.11h). There was also an interaction of participant group and visual articulation (β = 1.01, CI = 0.08, 1.93) such that American English participants were more likely to categorize visual articulation [ʋ]/ ‘w’ as /v/ for the Indian English speaker (Figure 2.11d). Tukey-adjusted pairwise comparisons revealed no effects of intervention in any of the subset conditions (Table 2.13).

Intervention	Visual Articulation	Model Talker	Group	Est/Lower CI/Upper CI
pre/post	[v]	AE	AE	-0.00176 -0.6655 0.658
pre/post	[w]	AE	AE	0.06507 -0.5520 0.671
pre/post	[ʋ] ‘v’	IE	AE	-0.04564 -0.7290 0.644
pre/post	[ʋ] ‘w’	IE	AE	0.13419 -0.5327 0.780
pre/post	[v]	AE	IE	-0.46535 -1.0728 0.188
pre/post	[w]	AE	IE	0.37745 -0.2843 1.041
pre/post	[ʋ] ‘v’	IE	IE	-0.00253 -0.6018 0.614
pre/post	[ʋ] ‘w’	IE	IE	-0.26234 -0.8725 0.387

Table 2.13: Pairwise comparisons for effects of intervention on illusory categorization in labial /i_i/ blocks.

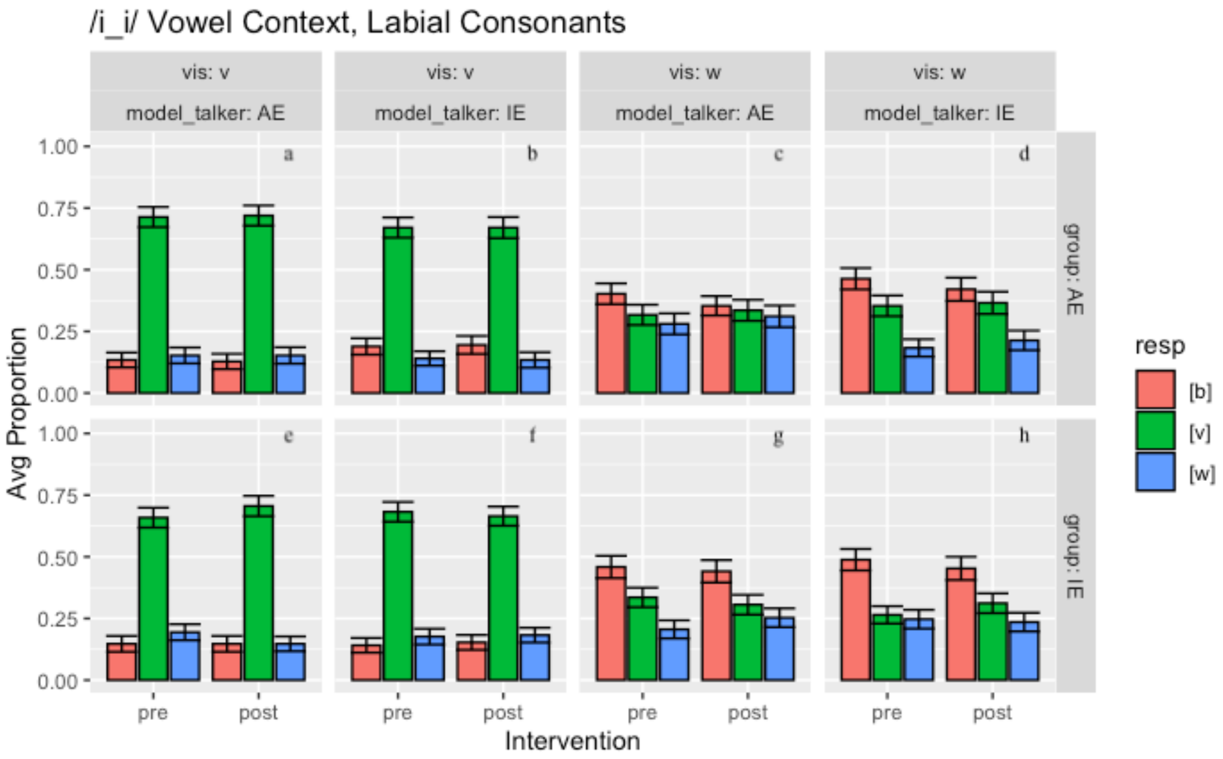

Figure 2.11: Rates of illusory categorization in the labial /i_i/ block. Each panel of the figure shows the proportion of illusory and veridical responses for a particular visual articulation (vis: v, w), model talker (speaker: AE, IE), participant group (group: AE, IE) condition and the standard error.

2.4 Discussion

2.4.1 General Discussion

This study examined the influence of sociophonetic experience on the categorization of illusory speech percepts by speakers of American English and Indian English. This was carried out by using a perceptual adaptation paradigm where participants categorized stimuli designed to elicit illusory percepts before and after gaining experience with the natural speech accent of two model talkers (American English & Indian English). The research questions at the center of this experiment were 1) Does the proportion of illusions experienced by participants change after gaining experience with the accent of the model talker and 2) Do illusory categorization rates change after participants gain experience with the accent of the model talkers?

With regard to the proportion of illusions changing as a function of experience, we see some evidence of this being the case but it is largely constrained to the [a] vowel context and to the visual articulations [g] and [w]/[ʋ] as opposed to [ð]/[d̤]and [v]/[ʋ]. Table 2.14 outlines the conditions where participants in each group shifted their illusion rate between pre- and post-intervention categorization tasks. One interpretation of the asymmetric patterns is that they may reflect the magnitude of visual cues available across conditions. The visual clarity of the American English [ð] and [v] articulations, for which visual cues play a key role in discriminating [v]/[ð] and [f]/[θ] in American English given their acoustic similarity, likely contributes to the patterns seen with [v] and [ð]. As a result, the [v] and [ð] stimuli may not be confusable enough, which would inhibit any effect that the intervention might have. However, in the case of the [g] and [w]/[ʋ] visual articulations, there did appear to be a somewhat consistent effect of intervention such that participants experienced more illusions with these stimuli after gaining experience with the speaking patterns of the model talkers. As for why these effects are constrained to the [a] vowel context, it may be the case that, due to the lower jaw position required for [a], the more extreme articulatory movement resulting in greater oral aperture yields visually transparent gestural paths into/out of the consonants that participants are tracking. Compare this with [i], which can be maintained with minimal oral aperture and no lower jaw movement.

Model Talker	Participant	Audio	Visual	Vowel
Within Dialect				
Indian English	Indian English	[b]	[ʋ]	a
Indian English	Indian English	[b]	[ḍ̤]	a
Indian English	Indian English	[b]	[g]	a
American English	American English	[b]	[w]	a
American English	American English	[b]	[g]	a
Across Dialect				
American English	Indian English	[b]	[g]	a
American English	Indian English	[b]	[g]	i
Indian English	American English	[b]	[ʋ]	a
Indian English	American English	[b]	[ʋ]	a
Indian English	American English	[b]	[g]	a
American English	American English	[b]	[g]	i

Table 2.14: Conditions for which there was an increase in the number of illusions between pre- and post-intervention.

With regard to the categorization trends, there was little evidence that listeners broadly shifted their categorization strategies as a result of the intervention. When participants did shift their categorization patterns, they did so in a way that seemed rooted in the experience they received from the model talkers. Table 2.15 below summarizes the conditions where participants in each group shifted their categorization strategy between pre- and post-intervention categorization tasks. Indian English participants showed a shift in categorization trends after intervention but only when perceiving the Indian English model talker. While this within dialect categorization shift was unexpected, all shifts reflected the veridical speech patterns that participants were exposed to in the intervention phase. Indian English participants, when responding to the Indian English model talker, showed post-intervention increases in /d/ categorizations when presented with a [ḍ̤] visual articulation (Figure 2.7f), and /v/ categorizations when presented with a [ʋ] visual articulation (Figure 2.9h). Interestingly, these shifts are closer to what was predicted for American English participants perceiving the speech of the Indian English model talker given their baseline productions.

Model Talker	Participant	Audio	Visual	Context	Change
Indian English	Indian English	[b]	[ḍ̤]	/i_i/	Increase in /d/ in post.
Indian English	Indian English	[b]	[ʋ]	/a_a/	Increase in /v/ in post.
American English	American English	[b]	[ð]	/i_i/	Increase in /ð/ in post.
American English	American English	[b]	[v]	/a_a/	Increase in /v/ in post.
American English	American English	[b]	[w]	/a_a/	Increase in /w/ in post.
Indian English	American English	[b]	[ʋ]	/a_a/	Increase in /v/ in post.

Table 2.15: Conditions for which there was a shift in categorization between pre- and post-intervention.

These trends, while reflective of the veridical speech of the Indian English model talker, could speculatively be due to the online presentation of the stimuli. Despite performing the task in India, the first screens that participants interacted with were consent forms, in English, noting that The University of Michigan was sponsoring the study that they were about to take part in. It could be the case that Indian English participants thought they were observing an Indian American model talker who could have an American English accent. Considering that participants received no experience with the veridical speech patterns of the model talkers before the pre-exposure task, if Indian English participants were assuming that the Indian English model talker was an Indian American, we would expect to see these exact categorization shifts outlined above. This is returned to in greater detail in Chapter 4.

As summarized in Table 2.15, only American English participants showed a categorization shift both across and within dialects. When responding to the American English model talker, American English participants showed increases in /ð/ when presented with a visual articulation of [ð] in the /i_i/ context (Figure 2.7a), as well as increases in /v/ (Figure 2.9a) and /w/ (Figure 2.9c) when presented [v] and [w] respectively in the /a_a/ context. In the case of [ð], the shift seen in categorization may be due to differences between vowel conditions. In the /a_a/ context, /ð/ responses appear to be approaching ceiling (Figure 2.5a) while in the /i_i/ context they appear to be around chance (Figure 2.7a). This difference in pre-intervention responses allows for the potential for shift, and considering that [i_i] vowel contexts are reported in the literature to yield greater /d/ responses (Green et al. 1998, Burnham & Dodd 2017), the within dialect shift seen here may be evidence that it was difficult for participants to make out the tongue tip of the model talker when they were producing [ð] and used the intervention experience to shift their categorization strategy in a speaker-specific way.

In the case of the labial consonants, American English participants again showed shifts in their categorization strategies and these shifts crucially extended across dialect boundaries. When responding to the American English model talker, American English participants showed an increase in /v/ responses when viewing [v] visual articulations (Figure 2.9a) and /w/ responses when viewing [w] visual articulations (Figure 2.9c) after intervention. When responding to the Indian English model talker, again after intervention, American English participants showed an increase in /v/ responses when viewing [ʋ] 'w' visual articulations. This pair of trends in opposite directions, crucially post-intervention, with visually articulated [w]/[ʋ] between the two model talkers provides the most compelling evidence that listeners are imputing accented characteristics onto the illusory percepts they're experiencing. This suggests that participants are learning about the production patterns of the model talkers and implementing this acquired phonological knowledge in the course of their illusory perception. In Figure 2.9c and Figure 2.9d it is clear that American English participants are largely failing to McGurk in the pre-intervention trials of the experiment as predicted. However, after intervention they not only appear to overcome this trend but also settle on different categorization strategies for each model talker. For the Indian English model talker, this strategy mirrors their experience with his natural speech patterns in the intervention phase – put another way, their categorization strategy more clearly reflects the phonology of the model talkers than that of the participant.

2.4.2 American English Participants and Illusory [w]

In §2.3.2 I suggested that because [w] has no burst or frication for the auditory [b] to map onto that we would expect that participants should simply fail to experience an illusion and report hearing [b]. However, American English participants were instead more likely to categorize illusory stimuli with visual /w/ in the /a_a/ context as /w/ post-intervention (Figure 6c. Thinking back to §1.1.2 and §1.3, illusions are intentionally constructed to be imperceptible from the union of their incongruent component parts, but their components can be perceived in isolation by ignoring or omitting one of the two competing signals. This is most easily demonstrated by having participants close their eyes or look off screen. In cases such as these, by ignoring the visual input, participants would simply report hearing a [b]. Given that illusory percepts are generally described to have less perceptual clarity than their veridical counterparts (Rosenblum, 2019), it could be the case here that American English listeners, upon experiencing some kind of perceptual event but without a characteristic clarity by which they can identify the percept are simply lipreading what they saw the American English model talker articulate. Put another way, they may be solving a difficult perceptual task by ignoring an 'opaque' perceptual unit and prioritizing what clarity remains in the visual signal. This is arguably distinct from these same listeners' approach with the Indian English model talker for whom they show the expected post-intervention increase in illusion rate (Figure 2.7d) and the expected shift in categorization – categorizing the Indian English model talker's [ʋ] articulation as /v/ (Figure 2.9d).

While these categorization patterns give some sense of how participants are using the veridical experience they gain with the model talkers in illusory test conditions, categorization is a broad strokes dependent variable that only reveals the final perceptual destination. In Chapter 3, I turn to a more gradient but implicit measure, speech shadowing, to understand how similar the phonetic productions of shadowed productions are to participant baselines.

Chapter 3 Shadowing of Illusory and Veridical Percepts

This study investigates the influence of sociophonetic experience on the shadowing of illusory and veridical speech percepts by speakers of American English and Indian English. As was the case in the previous experiment (Chapter 2) I again utilize the fact that American and Indian English differ from one another in aspects of their consonantal inventories (e.g. /d/, /ð/, /v/, /w/, /ʋ/) that are susceptible to the McGurk effect. Unlike Chapter 2, which provided a broad strokes measure of the final percept that participants experienced, in the following experiment described below, I use a shadowing paradigm to probe how a listener's acoustic production patterns shift when they're asked to repeat what a model talker has just said. Analysis of the sub-categorical details of shadowed productions provides a more granular measure than do perceptual categorizations of how listeners might be shifting their production targets to facilitate some form of perceptual adaptation towards the model talker of each dialect.

The research question being investigated in this experiment is whether participants who are tasked with shadowing illusory multisensory stimuli produced by model talkers from different English dialects show production patterns that mirror their own (the participants') baseline production or the model talker productions. Listeners employ adaptive perceptual strategies when perceiving accented speech and these strategies are often made manifest in the productions of listeners turned speakers when they are tasked with shadowing the accent of a model talker. As discussed in Chapter 1, in everyday speech settings, listeners not only rely on the acoustic information they perceive auditorily, but often, these information signals are reinforced with visual information as well. This experiment utilizes illusory percepts as a special stimulus case to investigate how shadowers extend their perceptual strategies from their veridical experiences to illusory conditions where the audio and visual information streams are comprised of conflicting signals. As noted throughout this dissertation, illusory conditions are being utilized in this work as a strategic tool to probe whether linguistically specified visual signals facilitate socially indexed perceptual adaptation without acoustic reinforcement. In this experiment I use acoustic measures (F2-F3 and F2n) to exploit the perception-production link inherent in shadowing to gain a finer grained understanding of the perceptual quality of the illusions that participants experience and to determine whether these percepts, despite their lack of acoustic reinforcement, contain accent qualities from the model talkers.

As noted in Section 1.1.3, while there is open debate regarding the exact computational primitives and perceptual completeness between illusory and veridical percepts, the visual and auditory streams appear to be treated similarly when perceiving speech and are broadly implicated in similar neuronal bases. Given these broad similarities, this experiment was designed with the two following assumptions: (1) that illusory shadowing and veridical shadowing utilize similar perceptual processes and (2) that the percepts that listeners experience are similar across veridical and illusory conditions. These assumptions are critical for ascribing differences between illusory and veridical shadowing to phonological knowledge as opposed to differences in processing or perceptual clarity.

Participants completed four phases in the experiment: training, baseline, intervention, and shadowing. In the shadowing phase, stimuli were either veridical or illusory on any given trial. This design allows for probing not only how American English and Indian English participants' baseline productions differ from shadowed productions, the traditional aim of speech shadowing (reflecting deviations from participant phonology), but it also allows for comparisons between shadows of model talker accents in illusory and veridical conditions – comparisons that should reflect differences in perceptual detail between congruent/incongruent model talker phonology. In this way, baseline and veridical stimuli serve as a continuum through which illusory shadowed productions, given their lack of compositional transparency, can be assessed as more participant or model talker like. Like with Experiment 2, the basic hypothesis that I pursue is that listeners will use all available meaningful information from an incoming signal, and as such, listeners will experience percepts that reflect their experience with the model talkers. Specifically, in baseline trials, given their lack of experience with the model talkers, participants are expected to use their natural speech patterns, which should broadly match the description of their dialect (American English, Indian English) within the literature. After experience with the model talkers, participants are predicted to shift their production strategies from their baselines towards the production targets of the model talkers. The rest of this chapter describes the details of the methodology (§ 3.1), predictions (§3.2), results (§ 3.3), and discussion (§ 3.4) for this experiment.

3.1 Methods

3.1.1 Participants

Forty-six participants were recruited between the University of Michigan (7 Indian English, 25 American English) and York University (14 Indian English) in Toronto, Ontario to participate in the shadowing experiment. York University was used as a secondary collection site due to the difficulty in recruiting Indian English speakers in Ann Arbor, MI and the large Indian immigrant community in the Greater Toronto Area. As with the categorization experiment in Chapter 2, participants self-identified as English speakers who had learned English no later than between the grades of K-12 in India or the United States of America. Due to technical issues, largely electrical interference within the recordings made at the University of Michigan, none of the Indian English participants and only 11 American English participants recorded at the University of Michigan were included for analysis. Twelve of the Indian English participants recorded at York University were included for analysis (2 participants removed due to file loss shortly after recording). Indian English participants reported speaking one or more of Hindi, Punjabi, Urdu, Malayalam, Bengali and Tamil in addition to English either at home as a child or currently at home. One Indian English participant reported speaking Bangla at home as a child and realized /ʋ/ without a burst (which would otherwise denote a [b] production) at a rate similar to other Indian English participants. In addition to monolingual American English participants, American English participants reported speaking one or more of Spanish, French, Russian, Hebrew, German, and Italian. One American English participant also reported signing American Sign Language. No participant self-reported a history of a speech or hearing disorder diagnosis, and all participants were either paid $20 USD (University of Michigan) or with a $20 CAD gift card (York University).

3.1.2 Stimuli

3.1.2.1 Audio Materials

In addition to the stimuli described in Section 2.1.2.1, audio stimuli were also made with the same American English and Indian English model talkers for veridical versions of the non-word sequences: [idi], [iði], [ivi], [iwi], [iḍ̤i], [iʋi], [ada], [aða], [ava] [awa], [aʋa], and [aḍ̤a]. These tokens were recorded during the same session as the [ibi] and [aba] sequences and followed the same recording protocols described in §2.1.2.1. Model talkers recorded the English non-words from a randomized list where they repeated each target item 10 times. Model talkers were instructed to produce the non-words with a trochaic stress pattern. From the 10 repetitions, the best production of each non-word, those free of noise or mispronunciations, was selected for inclusion. Audio recordings were made in a sound attenuated booth at the University of Michigan Phonetics Laboratory.

Like the [aba] and [ibi] tokens used in illusory audiovisual stimuli, tokens selected for the veridical set of audio portion of the audiovisual stimuli were also edited to reflect the average vowel and consonant durations of the American English and Indian English [aba] and [ibi] stimuli. This was done not only in an effort to neutralize speaker-specific artifacts, both socioindexical and otherwise, but also to ensure that veridical and illusory stimuli primarily only differed in the spectral information that the listeners were afforded. Despite inherent durational differences from manner and place of articulation in the natural productions of [b], [d], [ð] [ḍ̤], [v], [w] and [ʋ] the veridical set of consonants were edited to reflect the average duration of [ibi] and [aba] to ensure that durational cues would not confound the experiment. Table 3.1 below provides the durations for each consonant and vowel in the veridical stimulus recordings as well as the duration differences between the illusory audio (e.g. [ibi] and [aba]) and the veridical audio (e.g. [d], [ð], [d̤], [v], [w], [ʋ]).[16]

Indian English [ibi]	139	61	146	346
Indian English [idi]	133	69	144	346
Indian English [id̤i]	139	65	143	347
Indian English [iʋi] 'v'	132	62	147	341
Indian English [iʋi] 'w'	137	61	141	339
American English [ibi]	138	62	147	347
American English [idi]	134	60	146	340
American English [iði]	139	59	145	343
American English [ivi]	136	67	138	341
American English [iwi]	136	65	141	342
Indian English [aba]	153	62	107	322
Indian English [ada]	155	59	107	321
Indian English [ad̤a]	157	79	114	350
Indian English [aʋa] 'v'	159	53	107	319
Indian English [awa] 'w'	161	56	113	330
American English [aba]	153	63	108	324
American English [ada]	154	66	111	331
American English [aða]	155	63	108	326
American English [ava]	153	61	109	323
American English [awa]	154	60	108	322

Table 3.1: Durational values (in ms) for [b], [d], [ð], [d̤] [v], [w] and [ʋ] acoustic stimuli for the model talkers (American English, Indian English) across both vowel contexts (/i_i/, /a_a/).

[16] It should be noted that while [ibi] and [aba] are the audio for illusory audiovisual stimuli when paired with the visual articulations of [d], [ð], [d̤], [v], [ʋ], and [w], they also serve as audio in veridical audiovisual stimuli when they are paired with [b] visual articulations.

3.1.2.2 Video Materials

The video materials for this experiment were the same as those described in §2.1.2.2.

3.1.2.3 Audiovisual Materials

As described in §2.1.2.3, audio and video recordings were synced together using iMovie to make illusory and veridical audiovisual stimuli. For this experiment the set of veridical stimuli was expanded to include [d], [ð], [ḍ̤], [v], [w] and [ʋ]. All veridical audiovisual stimuli were created using the same protocols and procedures outlined in § 2.1.2.3. At the end of stimulus creation there were 16 additional stimuli to be paired with the original 20 illusory stimuli outlined in §2.1.2.3; Table 2.2. Table 3.2 below shows the configurations for the entire stimulus set with illusory stimuli from §2.1.2.3 highlighted in grey; non-highlighted stimuli are veridical.

Model Talker	Audio	Visual	Audiovisual	Audio	Visual	Audiovisual
AE	[aba]	[aba]	/aba/	[ibi]	[ibi]	/ibi/
AE	[aba]	[aga]	/ada/	[ibi]	[igi]	/idi/
AE	[aba]	[aða]	/aða/	[ibi]	[iði]	/iði/
AE	[aba]	[awa]	/aba/ or /awa/	[ibi]	[iwi]	/ibi/ or /iwi/
AE	[aba]	[ava]	/ava/	[ibi]	[ivi]	/ivi/
AE	[ada]	[ada]	/ada/	[idi]	[idi]	/idi/
AE	[aða]	[aða]	/aða/	[iði]	[iði]	/iði/
AE	[ava]	[ava]	/ava/	[ivi]	[ivi]	/ivi/
AE	[awa]	[awa]	/awa/	[iwi]	[iwi]	/iwi/
IE	[aba]	[aba]	/aba/	[ibi]	[ibi]	/ibi/
IE	[aba]	[aga]	/ada/	[ibi]	[igi]	/idi/
IE	[aba]	[aḍ̤a]	/aða/	[ibi]	[iḍ̤i]	/iði/
IE	[aba]	[aʋa]	/aba/ or /awa/	[ibi]	[iʋi]	/ibi/ or /iwi/
IE	[aba]	[aʋa]	/ava/	[ibi]	[iʋi]	/ivi/
IE	[ada]	[ada]	/ada/	[idi]	[idi]	/idi/
IE	[aḍ̤a]	[aḍ̤a]	/aḍ̤a/	[iḍ̤i]	[iḍ̤i]	/iḍ̤i/
IE	[aʋa]	[aʋa]	/ava/	[iʋi]	[iʋi]	/ivi/
IE	[aʋa]	[aʋa]	/awa/	[iʋi]	[iʋi]	/iwi/

Table 3.2: Complete set of stimuli used in the shadowing experiment. Audiovisual cells show predicted percepts that should be elicited from the conjunction of the audio and visual stimuli Audiovisual cells in grey are illusory pairings of VbV + visual articulation.

3.1.2.4 Intervention Video

The intervention video for this experiment was the same video described in § 2.2.4.

3.1.3 Procedure

Each participant was tested in a single experimental session either at the University of Michigan (11 American English participants) or at York University (12 Indian English participants). Participants completed the task in a sound attenuated booth, wearing AKG K271 MK II (University of Michigan) or Sennheiser HD 515 (York University) headphones, and the speech of each participant was digitally recorded in PRAAT (Boersma & Weenink, 2022) (University of Michigan) or Audacity (Mazzoni & Dannenburg, 2002) (York University) onto a computer using either a Røde PodMic (University of Michigan) or a Sony F-730 Dynamic microphone (York University). For participants at the University of Michigan, the author was the experimenter, for participants at York University, an undergraduate RA who was trained by the main author collected a majority of the data[17]. During the roughly 40-minute session, participants completed a speech shadowing task which consisted of a training portion followed by a baseline and shadowing portion which were mediated by the intervention video. The latter two portions, baseline and shadowing, were recorded for analysis. For the same reasons as those noted in § 2.1.3, there were no unimodal trials included in the shadowing experiment. At the end of all experimental tasks, participants again completed a language background questionnaire.

3.1.3.1 Speech Shadowing Task

The speech shadowing task was divided into four phases: 1) training, 2) baseline recording, 3) intervention, and 4) shadowed recording.

[17] The author collected data for three participants with the help of the undergraduate RA. These participants sessions served as training for the RA in addition to data collection.

Training: In training, participants were familiarized with the orthography that they'd be reading from the computer screen during the baseline recording portion of the experiment. Participants were trained by the experimenter on vowels and consonants separately with vowels proceeding consonants. For both the vowels and consonants, participants were read a script that described the need for a unique experimental orthography and participants were provided with examples of real English words that utilized the sounds they needed to link to the orthography. Table 3.3 below shows the sounds, experimental orthography, and English example words used for learning each of the target sounds. For example, when instructing participants how to produce /i/ (orthographically, "ii") participants were told, "When you see 'ii' in the experiment, it will be produced with the same vowel as in the word *trees* in English". A 5-vowel system was used in training and baseline to 1) acquire [u] productions from participants which was needed to calculate each participant's F2n and 2) to include distractor trials (e.g. [eɪ] and [o]) in baseline so that participants didn't become fixated on their [i] and [a] productions. After participants were taught how to produce each sound, they were tested by the experimenter with 10 random test trials (two presentations for each test vowel and consonant separately) of a VCV frame written in the experimental orthography (e.g. "iivii" [ivi] or "eibei" [eibei]). After participants successfully produced the target sound in the test trial they were moved onto the next random test trial by the experimenter. For training trials where the participant produced a sound incorrectly, participants were given multiple attempts to correct their production before being instructed what the correct sound was by the experimenter. After all test sounds has been correctly elicited using the experimental orthography in the training phase, the experimenter took a sound check to get the appropriate recording levels and left the booth to control the recording and experimental software from outside the booth.

Vowels	Orthography	Example Word	Consonants	Orthography	Example Word
/i/	ii	trees	/b/	b	beak
/a/	aa	box	/d/	d	door
/u/	uu	tube	/ð/	dh	this
/e/	ei	mail	/v/	v	veer
/o/	oe	hose	/w/	w	wilt

Table 3.3: Sounds, orthography, and example words from the training phase of the shadowing task

Baseline: In the baseline phase of the shadowing task, participants read the 25 (5 vowels x 5 consonants) English non-words in the experimental orthography that they had been trained on from a screen inside the sound booth. The participant's screen mirrored the content on the experimental software computer outside of the booth which was controlled by the experimenter. On each trial participants saw a fixation cross that drew their attention to the center of the screen and lasted 250ms after which they were presented with the English non-word experimental stimulus. As participants produced baseline tokens, the experimenter monitored their productions and advanced participants onto the next trial when a stimulus was correctly produced. When a stimulus was not correctly produced, the experimenter left the test stimulus on the screen and the participant tried again until correctly producing the stimulus word. Upon correctly producing the stimulus, participants were advanced to the next trial by the experimenter. In this way, the fixation cross at the beginning of a trial also served as a passive cue that they had produced a target stimulus correctly and had been moved onto the next trial. Baseline trials with errors were flagged by the experimenter and later, during Praat textgridding, only the correct final production was included for analysis. The baseline phase consisted of one fully randomized block of 125 trials (25 words x 5 repetitions) that included a timed 30 second break at the halfway point. During the break, a timer was projected onto the screen in front of the participant so that they would know when the baseline phase would restart.

Intervention video: At the end of all baseline trials, participants were given control of the experimental computer via a Bluetooth mouse and keyboard and instructed to continue to the intervention video. Participants watched the video (described in § 2.1.2.4), where they were exposed to the speech patterns of each of the model talkers. As described in § 2.1.3.1, participants were instructed to press play on the intervention video and to pay attention to the content. As was the case with the categorization experiment in Chapter 2, all participants passed the attention check trial.

Shadowing: After the intervention video, participants moved onto the shadowing phase of the experiment. In shadowing, unlike the baseline phase, participants controlled the pacing of the experiment. For each trial, a video would appear on the screen and the participant would press play. The video would play one of the audiovisual stimuli (veridical or illusory) and upon completion would advance to a screen that read "Please repeat what the speaker said". The screen for eliciting a response of what the participant heard stayed up for 2 seconds before moving onto the video screen for the next trial. Shadowing consisted of 200 trials (5 blocks x 40 trials) where every participant saw the illusory audiovisual stimuli (/d/, /ð/, /v/, /w/ x 2 vowels x 2 model talkers = 16), their veridical audiovisual counterparts (/d/, /ð/, /v/, /w/, x 2 vowels x 2 model talkers = 16), and two repetitions of the /b/ stimuli (/b/ x 2 reps x 2 vowels x 2 model talkers = 8; 16 + 16 + 8 = 40) in a given block. Each block was fully randomized and contained stimuli from both model talkers (Indian English and American English) and both vowel conditions (/a_a/ and /i_i/). At the end of each block, participants would receive a short break while the experimenter saved their sound file and started a new recording for the subsequent block.

3.1.4 Questionnaire

After the final shadowing block, participants were provided with nearly the same questionnaire as used in Experiment 1 (described in §2.1.3.2, Appendix C). The only change to the Experiment 2 questionnaire was the final question which omitted, “Did you use wired or Bluetooth headphones for this task?” and replaced it with “What did you think this experiment was about?”. This allowed for experimenters to debrief participants in a manner where they could probe the answers to this question further. American English responses largely centered around how “lip reading” facilitated speech perception, with and without reference to accents, or about perceiving the content of the nonsense words and test sounds. Indian English responses centered on the accents of Indian English speakers, how Indian English accents related to ‘English speaking’, and conceptions of ‘difference’ when learning English in India. As one Indian English participant put it succinctly, “i think this is about the accent.”

3.1.5 Acoustic Measures

In this experiment I employ two different dependent variable measures: F2-F3 for /b/, /d/, /ð/ stimuli and F2n for /v/ and /w/ stimuli. F2-F3 is the difference in Hz between the second and third formants of a given consonant. F2-F3 has been used in previous work that establishes distinct targets for retroflection in Indian English (Sirsa & Redford, 2013) as well as in general descriptions of retroflexed consonants in languages spoken in India (Ladefoged & Bhaskararao, 1983; Wiltshire, 2005; Wiltshire & Harnsberger, 2006, Wiltshire 2020). As Haman (2003) reviews, retroflexion (e.g. [ɖ]/[ʈ] in Indian English; high F2/low F3) is primarily cued by a lowering of F3 and asymmetrically shows extensive F3 lowering in VC transitions while CV transitions have much less extensive F3 lowering. This lowering is in conjunction with a high F2 which not only characterizes coronal articulations more generally but is also more susceptible to effects of vowel context than F3. Figure 3.1 below shows this asymmetry in the formant trajectories for [aɖa] produced by the Indian English model talker (elicited from an /ada/ prompt). Note the smaller difference between F2-F3 at the VC transition where F3 lowers dramatically going into the consonant closure (area between the two ovals) and nearly overlaps with F2. This can be compared with the CV transition coming out of the consonant closure, which shows a much larger F2-F3 difference in line with the asymmetry described by Haman (2003).

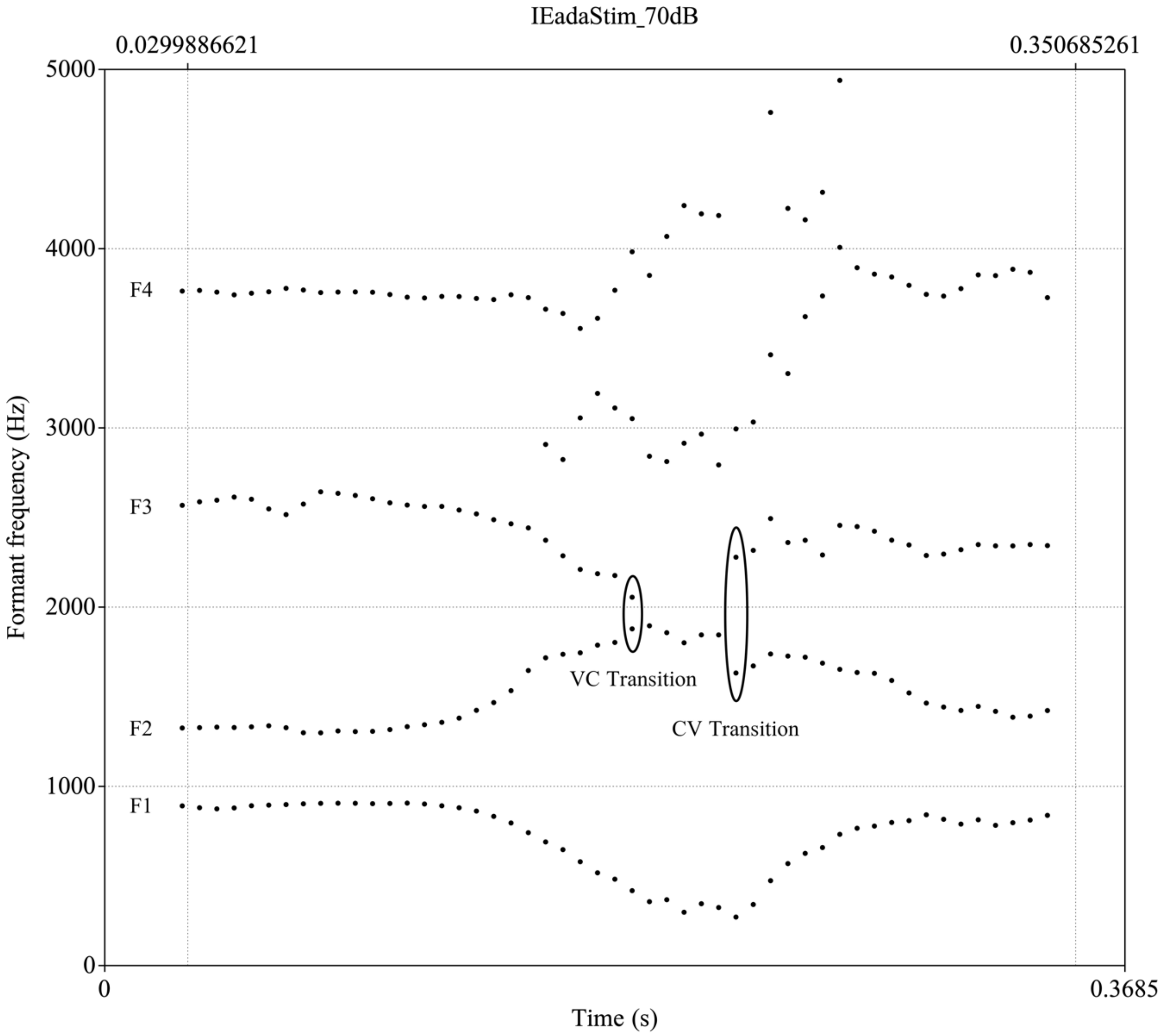

Figure 3.1: Formant trajectories for [aɖa] as produced by the Indian English model talker. VC and CV transition are marked by each ellipse with consonant closure occurring between the two ellipses. Note the F3 lowering at the VC transition for Indian English [aɖa].

[d] shows more symmetric F2-F3 effects. Figure 3.2 shows the formant trajectory of an [ada] sequence (elicited from an /ada/ prompt) produced by the American English speaker. As expected, F2 rises going into the coronal constriction, as does F3 (i.e., there is no evidence of retroflexion), yielding a comparatively large F2-F3 separation. Given these patterns, the smaller the F2-F3 difference, the more retroflexed or posterior a (shadowed) coronal production, and, in a similar vein, the larger the F2-F3 difference, the more /b/- (low F2/mid F3) or /d/-, /ð/- (high F2/mid F3) like a (shadowed) production. Measurements were taken both at the onset and

offset of consonant closure to understand participant production strategies throughout the entirety of the consonant.

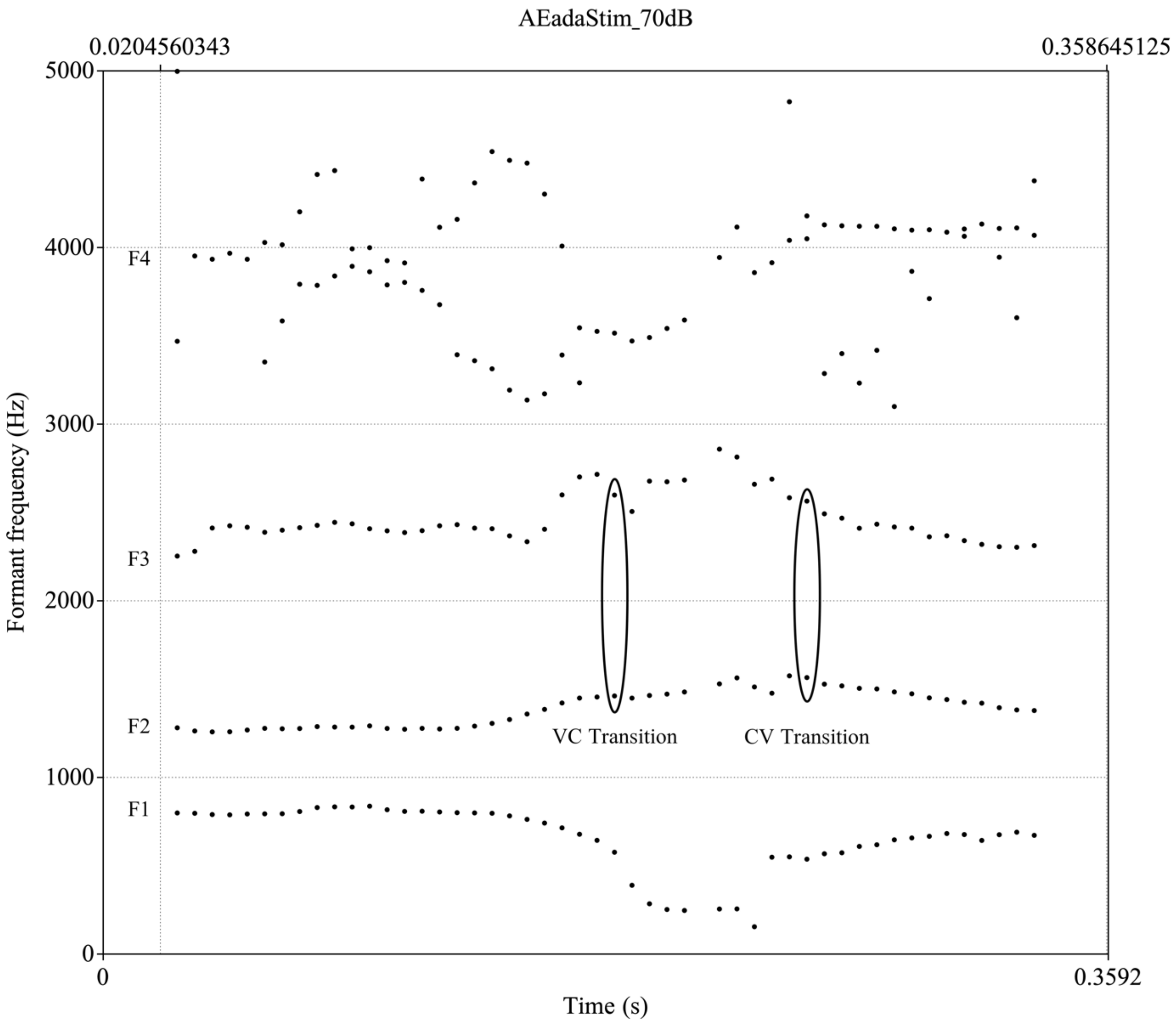

Figure 3.2: Formant trajectories for [ada] as produced by the American English model talker. VC and CV transition are marked by each ellipse with consonant closure occurring between the two ellipses. Note the lack of F3 lowering at VC transition for American English [ada].

F2n is a normalized F2 measure based on Fuchs (2019) that attempts to capture, “The normalized frequency of the second formant as a measure of lip-rounding and a greater velar constriction, where a lower F2n indicates more lip-rounding and a greater velar constriction” (Fuchs, 2019 pp.3). Fuchs (2019) argues that Indian English /v/ and /w/ appear to be in a state of near (but not complete) merger given that speakers do not produce differences in measures of frication, but do produce small differences, as compared to British English speakers, in F2 characteristics as measured by F2n. F2n is calculated as in Equation 1 where *$F2_n(w)$* is the normalized second formant of a phoneme and is equal to the inverse of the *MEAN F2* (grand mean of the second formant of all the /i/ and /u/ productions of a given participant) minus the F2 of a given phoneme divided by the *MEAN F2.*

$$F2_n(w) = -\frac{MEAN\ F2 - F2(w)}{MEAN\ F2}$$

Equation 3.1: Normalized F2 from Fuchs (2019)

In this experiment, F2n was measured at the midpoint of the consonant constrictions for all shadowed tokens that did not contain a stop burst. Consonants without a burst were selected for analysis because the presence of a burst suggests that there was no illusion and rather the participant responses were based on hearing acoustic [b]. Spectrograms and waveforms were visually inspected for a transient burst in every shadowed production from all participants. This guideline resulted in 75% (1396/1840) of all shadowed tokens being included for analysis with 29% (537/1840) coming from illusory stimuli and 46% (859/1840) coming from veridical stimuli.

3.1.6 Statistical Analysis

The research question being investigated in this experiment is whether participants who are tasked with shadowing veridical and illusory multisensory stimuli produced by model talkers from different English dialects show production patterns that mirror their own (the participants') baseline production strategy. To investigate this question statistically, Linear

Mixed-Effect models were run using the `lme4` package in R to model F2-F3 and F2n in participant responses' to coronal and labial stimuli, respectively. Separate statistical models were run for each place of articulation (coronal: /b/, /d/, /ð/ and labial: /v/, /w/) given the different dependent variables for each stimulus type, and for each vowel condition given the effects of vowel context on the dependent variables. Coronal models were further divided into models that measured F2-F3 at consonant onset and consonant offset given the asymmetric distribution of cues found in retroflexes. This results in four coronal models (/a_a/ context at consonant onset, /a_a/ context at consonant offset, /i_i/ context at consonant onset, and /i_i/ context at consonant offset) and two labial models (/a_a/ context and /i_i/ context).

For coronal models, the dependent variable was F2-F3. The effects of experimental phase (baseline, shadowing), group (American English, Indian English), model talker (American English, Indian English, self), visual articulation ([b], [d], [ð]), and stimulus type (baseline, veridical, illusory), their interactions, and a random intercept for participant were included in the model. For labial models, the dependent variable was F2n as computed in Equation 1 above. Like the coronal models, the effects of phase (baseline, shadowing), group (American English, Indian English), model talker (American English, Indian English, self), visual articulation (/v/ and /w/), and stimulus type (baseline, veridical, illusory), their interactions, and a random intercept for participant were included in the model. Tukey-adjusted pairwise comparisons were also computed using the `emmeans` package in R to better understand model results.

3.2 Predictions

3.2.1 American English Baselines

As noted in §2.2.1, American English participants are expected to come to the task with separate phonemic categories for /d/, /ð/, /v/, and /w/, and in the case of /d/, /ð/, and /v/ there are clear visual articulations within American English that facilitate illusory assignment of the burst of [b] to one of these three categories. Anticipating the results of the baseline measures in §3.3,[18] American English participants reliably produced significant differences in F2-F3 between /b/ and /d/ given their differences in place of articulation. However, they did not produce an F2-F3 difference between /b/ and /ð/. American English participants also produced significant differences in F2n between /v/ and /w/ in their baseline recordings.

3.2.2 American English Shadowing Predictions

Under the assumptions laid out above in the introduction, I predict that American English participants will not only McGurk for illusory stimuli, but that the F2-F3 and F2n values for their within dialect illusory shadows will not differ from their within dialect veridical shadows or their baseline productions. That is, shadowing will operate in a uniform fashion within dialect and participants will only show a shift in their coronal and labial productions reflected in the F2-F3 and F2n dependent measures, respectively, when shadowing across dialects.

For the coronals, I predict that American English participants will show a smaller F2-F3 difference for /d/, relative to their baseline F2-F3 measures for /d/, when shadowing the Indian

[18] Baseline F2-F3 measures for non-retroflex coronals /ð/ and /d/ did not differ under any condition (American English and Indian English model talkers, /a_a/ and /i_i/ vowel conditions, and consonant onset/offset). As such, no predictions are being made for that comparison.

English model talker reflecting the retroflexion cued via acoustic reinforcement (veridical) and listener experience with that retroflexion (illusory) in the Indian English model talker's productions. These differences should hold in both vowel conditions (/a_a/ and /i_i/), though in the /i_i/ condition the effect may be diminished given the high F2 of /i/ and the susceptibility of F2 in coronal consonant transitions to coarticulatory vowel effects. For /v/ and /w/, I predict that American English participants will maintain their baseline F2n differences between /v/ and /w/ when shadowing within dialect. However, I expect that American English participants will shift their production strategies such that there will be no statistical difference in F2n between /v/ and /w/ when shadowing the Indian English model talker. This would reflect the merger present in the speech of the Indian English model talker. Like with F2-F3, I expect that the difference in F2n will hold in both veridical and illusory shadowing conditions as well as in both vowel contexts. Table 3.4 summarizes these predictions for American English participants across all relevant contrasts[19].

[19] While American English participants could produce more retroflexed coronals while shadowing the Indian English model talker in /ð/ conditions, because there is no difference between /d/ and /ð/ in the F2-F3 measures I've omitted any predictions. Likewise, I am predicting no change in /ð/ because results would be undecipherable from /b/ given the lack of a difference in the baseline results.

American English Participants	Acoustic Measure	Baseline (Actual Results)	Predicted Veridical Results (Within)	Predicted Veridical Results (Across)	Predicted Illusory Results (Within)	Predicted Illusory Results (Across)
/d/	F2-F3	Large F2-F3 difference (anterior production)	Same as baseline	Smaller F2-F3 difference compared to baseline /d/ (posterior production)	Same as baseline	Smaller F2-F3 difference compared to baseline /d/ (posterior production)
/ð/	F2-F3	No difference between /b/ and /ð/[20]	Same as baseline	Same as baseline	Same as baseline	Same as baseline
/v/ v. /w/	F2n	Difference between /v/ and /w/	Same as baseline	No difference in F2n for /v/ and /w/	Same as baseline	No difference in F2n for /v/ and /w/

Table 3.4: American English baseline results and predictions for F2-F3 and F2n in within and across dialect conditions.

3.2.3 Indian English Baselines

Unlike American English participants, Indian English participants are expected to come to the task with phonemic categories for /d/ and /ʋ/. Anticipating the results of the baseline measures for Indian English in §3.3., participants reliably produced significant differences in F2-F3 between /b/ and /d/ and /b/ and /ð/. In line with the descriptions of Indian English

[20] The lack of difference between /b/ and /ð/ suggests that F2-F3 is not an interpretable dependent measure for assessing differences between /b/ and /ð/ for American English. As such, no predictions are made with regard to shifts in /ð/.

(Gargesh 2008, Sailaja 2012, Fuchs 2019, Wiltshire 2020) Indian English participants did not show a difference in F2n between /v/ and /w/ productions in baseline.

3.2.4 Indian English Shadowing Predictions

I predict that Indian English participants will experience McGurk effects for the illusory stimuli and that their F2-F3 and F2n values for within dialect illusory shadows will not differ from within dialect veridical shadowed productions or their baseline productions. Like the American English participants, Indian English participants should also shadow in a uniform fashion within dialect and should only show a shift in their productions reflected in the (coronal) F2-F3 and (labial) F2n dependent measures when shadowing across dialect.

Across dialect, I predict that, relative to their baselines, Indian English participants will show a larger F2-F3 difference when shadowing American English /d/ and /ð/. In the case of /d/ this is thought to reflect the acoustic landmarks (veridical) and the participant's linguistic experience (illusory). In the case of /ð/ this would reflect not only the production landmarks of the American English model talker but also the suppression of the substitution pattern found in Indian English participant baseline productions. As noted above, F2-F3 effects may be diminished in the /i_i/ vowel condition given the high frequency F2 of /i/.

In the case of /v/ and /w/, although Indian English participants' baseline productions do not show a difference in F2n between /v/ and /w/, I predict that a statistical difference will emerge when shadowing the American English talker. This shift would reflect the split present in the speech of the American English model talker (veridical) and the participant's linguistic experience (illusory). As was the case with the American English predictions laid out above, I expect that these effects will persist across both vowel conditions. Table 3.5 summarizes these predictions for Indian English participants across all relevant contrasts

Indian English Participants	**Acoustic Measure**	**Baseline (Actual Results)**	**Predicted Veridical Results (Within)**	**Predicted Veridical Results (Across)**	**Predicted Illusory Results (Within)**	**Predicted Illusory Results (Across)**
/d/	F2-F3	Small F2-F3 difference (posterior production	Same as baseline	Larger F2-F3 difference compared to baseline (anterior production)	Same as baseline	Larger F2-F3 difference compared to baseline (anterior production)
/ð/	F2-F3	Small F2-F3 difference (posterior production)	Same as baseline	Larger F2-F3 difference compared to baseline (anterior production)	Same as baseline	Larger F2-F3 difference to baseline (anterior production)
/v/ v. /w/	F2n	No difference between /v/ and /w/	Same as baseline	Difference between /v/ and /w/	No Change	Difference between /v/ and /w/

Table 3.5: Indian English predictions for F2-F3 and F2n in within and across dialect conditions.

3.3 Results

3.3.1 Results for coronals, /a_a/ context, at consonant onset

Figure 3.3 below graphically represents the results of the F2-F3 model for the /a_a/ vowel context at consonant onset (VC). The model revealed effects of [d] visual articulation ($\beta = 274.20$, $t = 4.618$, $p = < .0001$) and [ð]/[ḍ̤] visual articulation ($\beta = 167.56$, $t = 2.821$, $p = .00485$); these main effects were mediated by two- and three-way interactions. There were significant two-way interactions between group and visual articulation for both [d] (β = 229.363, t =2.852, p =.00440; Figure 3.3a vs. 3.3c vs. 3.3f) and [ð]/[ḍ̤] (β = 330.554, t = 4.094, p <.001) suggesting that, relative to American English participants, Indian English participants showed a greater F2-F3 difference between [b] and [d] and between [b] and [ð]/[ḍ̤] productions. There were also three-way interactions among (i) phase, group, and visual articulation /ð/ (β = -273.151, t=2.576, p=0.01008) such that Indian English participants showed (relative to American English participants) a larger difference in F2-F3 during shadowing [ð]/[ḍ̤] (a more anterior production) and (ii) group, stimulus type, and visual articulation [d] (β=268.795, t=2.360, p=0.01840) suggesting that Indian English participants showed a smaller difference in F2-F3 when shadowing veridical [d] (a more posterior production).

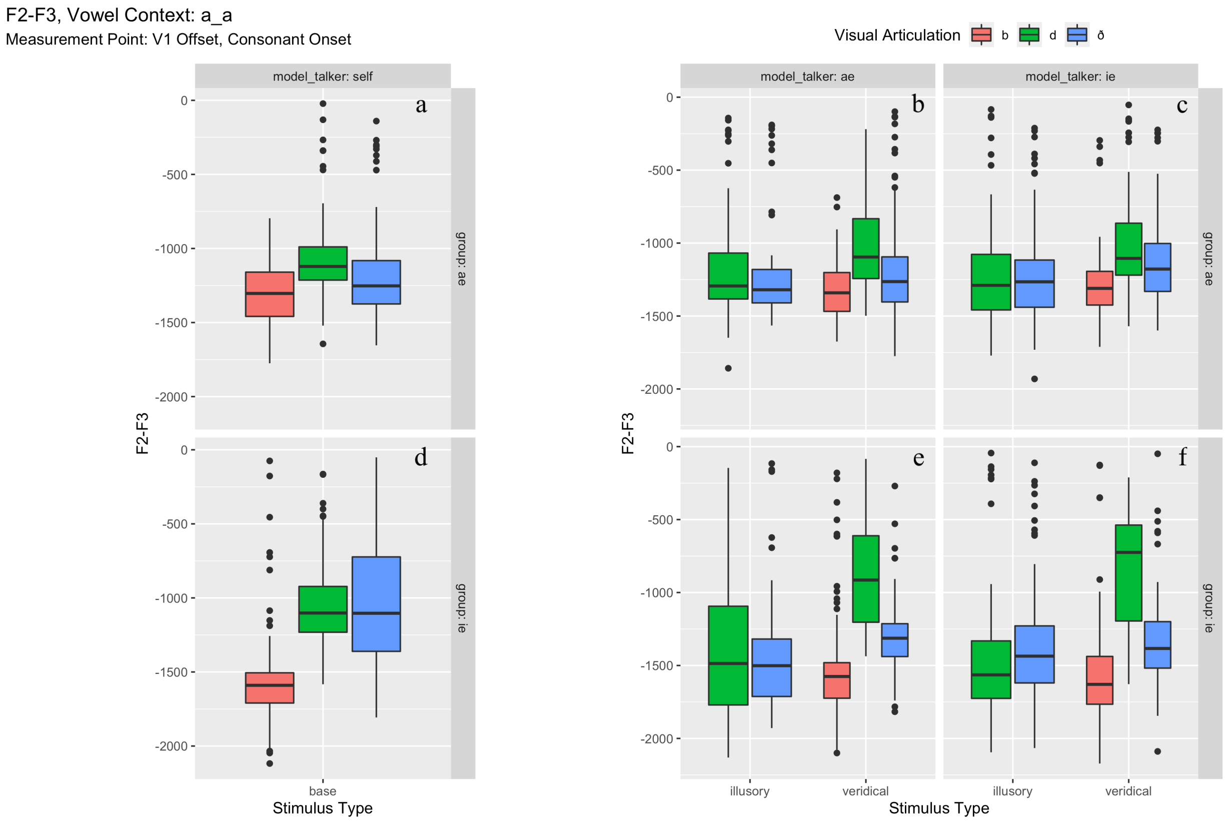

Figure 3.3: Boxplot of F2-F3 by stimulus type (baseline, illusory, veridical) for the /a_a/ context measured at consonant onset (VC). A higher value in F2-F3 denotes a more posterior production. Top row: American English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English). Bottom row: Indian English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English).

Tukey-adjusted pairwise comparisons were conducted to investigate whether there wasan effect of visual articulation between baseline, veridical and illusory stimulus types. Results of these comparisons are outlined below for American English participants (Table 3.6) and Indian English participants (Table 3.7). For American English participants there were no significant within-dialect or across-dialect shifts for any of the pairwise comparisons. As such, the precise production strategies of American English participants are unclear. While they may be using their baseline production strategies in veridical and illusory shadowing with both dialect groups, it also may be the case that F2-F3 isn't sensitive enough to reveal within or across category shifts for American English participants' productions.

For Indian English participants (Table 3.7), there was, as expected, a baseline difference between /b/ and /d/ and between /b/ and /ð/ but not between baseline /d/ and /ð/ (Figure 3.3d). Unexpectedly, Indian English participants showed evidence of a shift in not only the across, but also the within dialect conditions. For both model talkers, Indian English participants' illusory productions of /d/ and /ð/ had a larger F2-F3 difference (i.e., a more anterior articulation) than was found for their baseline productions (Figure 3.3e and Figure 3.3f when compared to Figure 3.3d). There was also an effect of stimulus type, such that illusory and veridical productions of /d/ (Figure 3.3e and 3.3f) were different from each other, which, when paired with the lack of a difference between baseline and veridical /d/ suggests—contrary to predictions—that veridical productions of /d/ were reflective of participant baselines while illusions were not. This is different from /ð/, which showed a F2-F3 difference between baseline and illusory (Figure 3.3d vs. Figure 3.3e and 3.3f) and baseline and veridical productions (again, Figure 3.3d vs. Figure 3.3e and 3.3f) but no F2-F3 difference between illusory and veridical conditions (Figure 3.3e and Figure 3.3f).

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /b/ vs. Baseline /d/	*-274.32*	*59.4*	*-4.618*	*0.0012* *
Baseline /b/ vs. Baseline /ð/	-167.56	59.4	-2.821	0.4591
Baseline /d/ vs. Baseline /ð/	106.76	59.4	1.797	0.9893
Within Dialect (AE Model, AE Participant), Vowel Context: a_a, C Onset				
Baseline /d/ vs. Illusory /d/	102.66	59.4	1.728	0.9937
Baseline /d/ vs. Veridical /d/	-45.02	59.4	-0.758	1
Illusory /d/ vs. Veridical /d/	-147.68	59.4	-2.486	0.7251
Baseline /ð/ vs. Illusory /ð/	-13.24	59.4	-0.223	1
Baseline /ð/ vs. Veridical /ð/	*43.43*	*59.7*	*0.727*	*1*
Illusory /ð/ vs. Veridical /ð/	*-56.67*	*59.7*	*-0.949*	*1*
Across Dialect (IE Model Talker, AE Participant), Vowel Context: a_a, C Onset				
Baseline /d/ vs. Illusory /d/	143.58	59.4	2.417	0.7738
Baseline /d/ vs. Veridical /d/	-60.48	59.4	-1.018	1
Illusory /d/ vs. Veridical /d/	-204.06	59.4	-3.435	0.1089
Baseline /ð/ vs. Illusory /ð/	-42.32	59.4	-0.712	1
Baseline /ð/ vs. Veridical /ð/	32.3	59.4	0.544	1
Illusory /ð/ vs. Veridical /ð/	-74.62	59.4	-1.256	1

Table 3.6: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant onset in the /a_a/ context for American English participants. Comparisons for which there is a significant effect are starred. Comparisons that are consistent with predictions are in italics.

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /b/ vs. Baseline /d/	*-503.68*	*54.2*	*-9.289*	*<.0001**
Baseline /b/ vs. Baseline /ð/	*-489.11*	*54.7*	*-9.107*	*<.0001 **
Baseline /d/ vs. Baseline /ð/	5.57	54.7	0.102	1
Within Dialect (IE Model, IE Participant), Vowel Context: a_a, C Onset				
Baseline /d/ vs. Illusory /d/	403.72	54.2	7.445	<.0001 *
Baseline /d/ vs. Veridical /d/	-198.77	54.2	-3.666	0.0532
Illusory /d/ vs. Veridical /d/	-602.48	54.2	-11.111	<.0001 *
Baseline /ð/ vs. Illusory /ð/	302.78	54.7	5.536	<.0001 *
Baseline /ð/ vs. Veridical /ð/	325.3	54.7	5.947	<.0001 *
Illusory /ð/ vs. Veridical /ð/	-22.52	54.2	-0.415	1
Across Dialect (AE Model Talker, IE Participant), Vowel Context: a_a, C Onset				
Baseline /d/ vs. Illusory /d/	*357.83*	*54.2*	*6.599*	*<.0001 **
Baseline /d/ vs. Veridical /d/	-120.28	54.2	-2.218	0.8875
Illusory /d/ vs. Veridical /d/	*-478.12*	*54.2*	*-8.817*	*<.0001 **
Baseline /ð/ vs. Illusory /ð/	*285.53*	*54.7*	*5.22*	*0.0001 **
Baseline /ð/ vs. Veridical /ð/	*403.85*	*54.7*	*7.384*	*<.0001 **
Illusory /ð/ vs. Veridical /ð/	-118.32	54.2	-2.182	0.9034

Table 3.7: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant onset in the /a_a / context for Indian English participants. Comparisons for which there is a significant effect are starred. Comparisons that are consistent with predictions are in italics.

3.3.2 Results for coronals, /a_a/ context, at consonant offset

Figure 3.4 shows the results of the F2-F3 model for the /a_a/ context at consonant offset (CV). The model revealed main effects of group (β = -226.097, t = -2.728, p = 0.00792) and of the [d] (β = 282.560, t = 4.615, p < .0001) and [ð]/[d̪] (β = 161.200, t = 2.633, p = 0.008) visual articulations. Both of these visual articulations also entered into a two-way interaction with group ([d]: β = 210.390, t = 2.538 p = 0.01118; [ð]/[d̪]: β = 234.746, t = 2.821, p = 0.00481) such that the F2-F3 difference for these productions was slightly smaller for the Indian English than for the American English participants (Figure 3.4a-c & 3.4d-f). This pattern is consistent with more backed and retroflexed productions for the Indian English group.

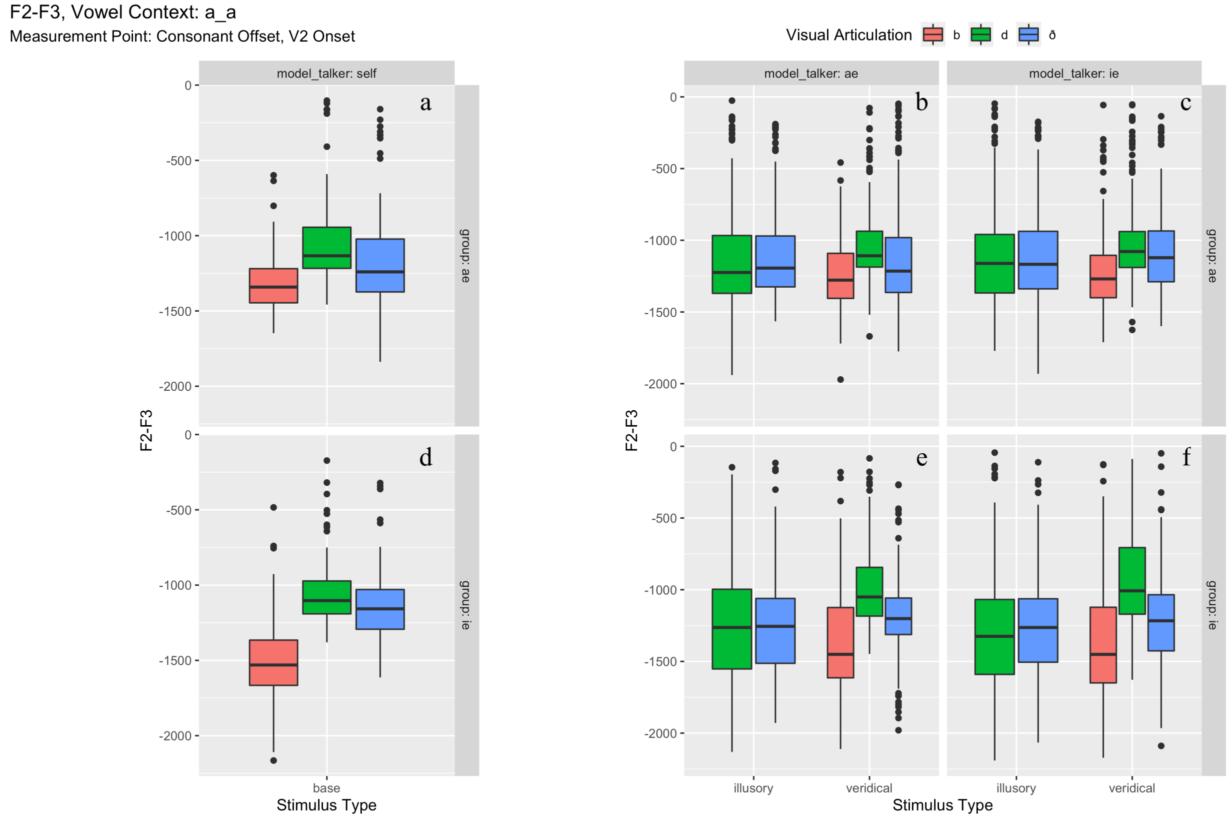

Figure 3.4: Box plot of F2-F3 by stimulus type (baseline, illusory, veridical) for the /a_a/ context measured at consonant offset (CV). Top row: American English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English). Bottom row: Indian English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English).

Tukey-adjusted pairwise comparisons were conducted to investigate whether there was an effect of visual articulation between baseline, veridical and illusory stimulus types. Tables 3.8 and 3.9 give the results of these comparisons for American English and Indian English participants, respectively. For American English participants (Table 3.8), there was effect of visual articulation in the baseline conditions such that the difference in F2-F3 was larger for the [b] than for the [d] visual articulations (Figure 3.4a). However, aside from that, there were no other effects for American English participants in either of the stimulus conditions (illusory/veridical) for either model talker. While this was unexpected, this pattern matches what was seen at the vowel offset/consonant onset in Table 3.5 above. For Indian English participants (Table 3.9), there was an effect of visual articulation in the baseline conditions for both [d] and [ð]/[ḍ̤] as expected (Figure 3.4d). Again, as was found for the measurements at vowel offset/consonant offset in Table 3.7, Indian English participants unexpectedly showed evidence of a shift in both the within and across dialect conditions. Indian English shadowers showed a difference between their baseline and illusory productions for [d] (compare Figure 3.4d to 3.4e and 3.4f) as well as a difference between their illusory and veridical productions of [d] (same figure panel comparison).

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /b/ vs. Baseline /d/	_-282.56_	_61.2_	_-4.615_	_0.0011 *_
Baseline /b/ vs. Baseline /ð/	-161.2	61.2	-2.633	0.6109
Baseline /d/ vs. Baseline /ð/	121.36	61.2	1.982	0.9648
Within Dialect (AE Model, AE Participant), Vowel Context: a_a, C Offset				
Baseline /d/ vs. Illusory /d/	107.947	50	2.159	0.9131
Baseline /d/ vs. Veridical /d/	16.327	50	0.327	1
Illusory /d/ vs. Veridical /d/	-91.62	35.3	-2.592	0.6439
Baseline /ð/ vs. Illusory /ð/	-37.527	50	-0.751	1
Baseline /ð/ vs. Veridical /ð/	-35.992	50	-0.718	1
Illusory /ð/ vs. Veridical /ð/	-1.534	35.5	-0.043	1
Across Dialect (IE Model Talker, AE Participant), Vowel Context: a_a, C Offset				
Baseline /d/ vs. Illusory /d/	77.76	50	1.556	0.9987
Baseline /d/ vs. Veridical /d/	-18.507	50	0.37	1
Illusory /d/ vs. Veridical /d/	-96.267	35.3	-2.723	0.5371
Baseline /ð/ vs. Illusory /ð/	-68.013	50	-1.361	0.9999
Baseline /ð/ vs. Veridical /ð/	-45.853	50	-0.917	1
Illusory /ð/ vs. Veridical /ð/	-22.16	35.3	-0.627	1

Table 3.8: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant offset in the /a_a/ context for American English participants. Comparisons for which there is a significant effect are highlighted in yellow. Comparisons that are consistent with predictions are in italics.

Contrast	Estimate	Standard Error	z.ratio	p.value
Baseline /b/ vs. Baseline /d/	*-492.95*	*55.9*	*-8.82*	*<.0001 **
Baseline /b/ vs. Baseline /ð/	*-395.947*	*56.4*	*-7.024*	*<.0001 **
Baseline /d/ vs. Baseline /ð/	97.004	56.4	1.721	0.9942
Within Dialect (IE Model, IE Participant), Vowel Context: a_a, C Offset				
Baseline /d/ vs. Illusory /d/	283.667	45.6	6.126	<.0001 *
Baseline /d/ vs. Veridical /d/	-95.178	45.6	-2.086	0.9387
Illusory /d/ vs. Veridical /d/	-378.844	32.3	-11.741	<.0001 *
Baseline /ð/ vs. Illusory /ð/	84.458	46.2	1.827	0.987
Baseline /ð/ vs. Veridical /ð/	129.48	46.2	2.801	0.4739
Illusory /ð/ vs. Veridical /ð/	-45.022	32.3	-1.395	0.9998
Across Dialect (AE Model Talker, IE Participant), Vowel Context: a_a, C Offset				
Baseline /d/ vs. Illusory /d/	*232.878*	*45.6*	*5.103*	*0.0001 **
Baseline /d/ vs. Veridical /d/	-47.094	45.6	-1.032	1
Illusory /d/ vs. Veridical /d/	*-279.972*	*32.3*	*-8.667*	*<.0001 **
Baseline /ð/ vs. Illusory /ð/	61.152	46.2	1.323	0.9999
Baseline /ð/ vs. Veridical /ð/	128.791	46.2	2.786	0.4859
Illusory /ð/ vs. Veridical /ð/	157.668	67.7	2.33	0.8297

Table 3.9: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant offset in /a_a/ context for Indian English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

3.3.3 Results for coronals, /i_i/ context, at consonant onset

Figure 3.5 shows the results of the F2-F3 model for the /i_i/ vowel context at consonant onset (VC). The model revealed a main effect of phase (β = 135.491, t = 3.511, p = .0004). Phase entered into a two-way interaction with group (β = -187.2338, t = -3.597, p = .0003) and three-way interactions with group and visual articulation for [d] (β = 249.733, t = 3.101, p = .001) and /ð/ (β = 233.4605, t = 2.909, p = .003) such that that Indian English participants produced [d] and [ð]/[d̤] with a smaller difference in F2-F3 (more posterior) during shadowing than in baseline (Figure 3.5.d compared to 3.5e and 3.5f). It is likely the case that the vowel environment obscured many differences in F2-F3 given the high F2 of [i] and F2's susceptibility to vowel environment effects.

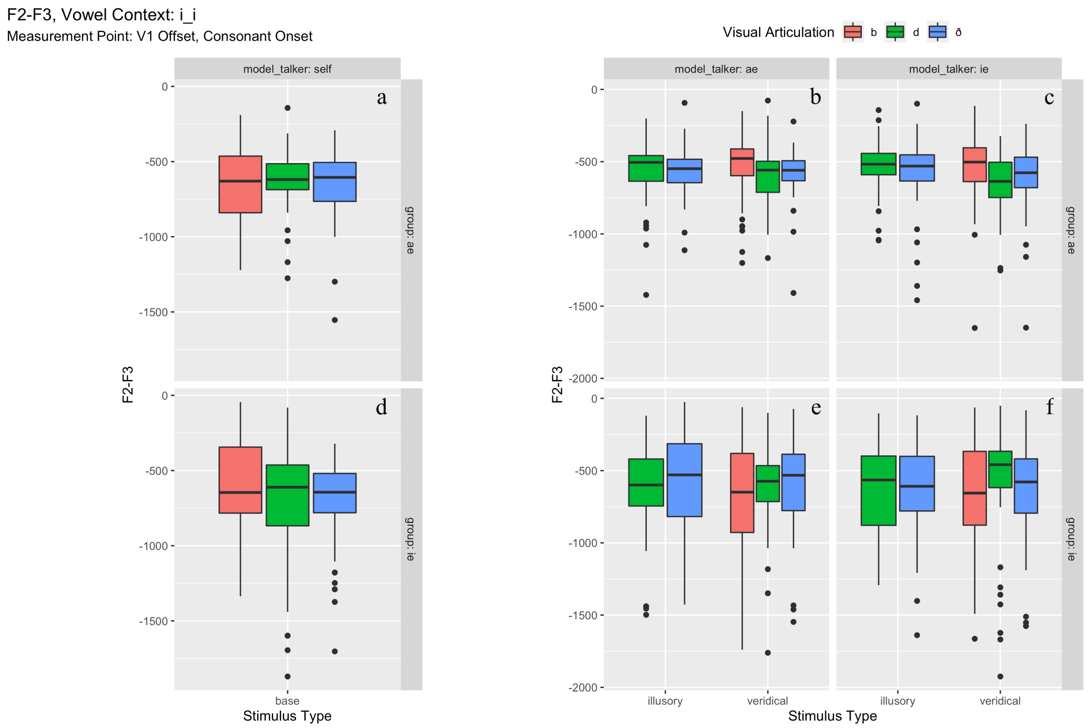

Figure 3.5: Box plot of F2-F3 by stimulus type (baseline, illusory, veridical) for the /i_i/ context measured at consonant onset. Top row: American English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English). Bottom row: Indian English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English).

Tukey-adjusted pairwise comparisons were conducted and no effects were found for American English (Table 3.10) or Indian English (Table 3.11) participants in any condition.

Contrast	Estimate	Standard Error	z.ratio	p.value
Baseline /b/ vs. Baseline /d/	-26.476	45.1	-0.587	1
Baseline /b/ vs. Baseline /ð/	2.94	45.1	0.065	1
Baseline /d/ vs. Baseline /ð/	29.416	45.1	0.652	1
Within Dialect (AE Model, AE Participant), Vowel Context: i_i, C Onset				
Baseline /d/ vs. Illusory /d/	-64.404	45.1	-1.428	0.9997
Baseline /d/ vs. Veridical /d/	-25.994	45.1	-0.575	1
Illusory /d/ vs. Veridical /d/	38.46	45.1	0.853	1
Baseline /ð/ vs. Illusory /ð/	-95.64	45.1	-2.12	0.9275
Baseline /ð/ vs. Veridical /ð/	-77.94	45.1	-1.728	0.9939
Illusory /ð/ vs. Veridical /ð/	17.7	45.1	0.392	1
Across Dialect (IE Model Talker, AE Participant), Vowel Context: i_i, C Onset				
Baseline /d/ vs. Illusory /d/	-93.244	45.1	-2.067	0.9442
Baseline /d/ vs. Veridical /d/	23.356	45.1	0.518	1
Illusory /d/ vs. Veridical /d/	116.6	45.1	2.585	0.6496
Baseline /ð/ vs. Illusory /ð/	-81.54	45.1	-1.808	0.9887
Baseline /ð/ vs. Veridical /ð/	-46.014	45.3	-1.015	1
Illusory /ð/ vs. Veridical /ð/	35.526	45.3	0.784	1

Table 3.10: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant onset in /i_i/ context for American English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

Contrast	Estimate	Standard Error	z.ratio	p.value
Baseline /b/ vs. Baseline /d/	88.967	41.2	2.16	0.913
Baseline /b/ vs. Baseline /ð/	93.767	41	2.286	0.8543
Baseline /d/ vs. Baseline /ð/	4.8	41.4	0.116	1
Within Dialect (IE Model, IE Participant), Vowel Context: i_i, C Onset				
Baseline /d/ vs. Illusory /d/	-56.091	41.5	-1.351	0.999
Baseline /d/ vs. Veridical /d/	-131.4	41.4	-3.177	0.2178
Illusory /d/ vs. Veridical /d/	-75.309	41.4	-1.821	0.9873
Baseline /ð/ vs. Illusory /ð/	-79.417	41.2	-1.929	0.974
Baseline /ð/ vs. Veridical /ð/	-71.183	41.2	-1.729	0.9937
Illusory /ð/ vs. Veridical /ð/	8.233	41.2	0.2	1
Across Dialect (AE Model Talker, IE Participant), Vowel Context: i_i, C Onset				
Baseline /d/ vs. Illusory /d/	-74.433	41.4	-1.8	0.9891
Baseline /d/ vs. Veridical /d/	-87.883	41.4	-2.125	0.9251
Illusory /d/ vs. Veridical /d/	-13.45	41.2	-0.327	1
Baseline /ð/ vs. Illusory /ð/	-141.633	41.2	-3.44	0.1075
Baseline /ð/ vs. Veridical /ð/	-124.167	41.2	-3.015	0.3156
Illusory /ð/ vs. Veridical /ð/	-44.75	41.2	-1.087	1

Table 3.11: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant onset in /i_i/ context for Indian English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

3.3.4 Results for coronals, /i_i/ context, at consonant offset

Figure 3.6 gives the results of the F2-F3 model for the /i_i/ vowel context at consonant offset/V2 onset. The model revealed a main effect of phase (β = -126.358, t = -2.876, p = .004) such that, relative to baseline, shadowed productions showed a slightly but significantly greater difference in F2-F3 (compare left panels of Fig. 3.6 to right and center panels). There was also a two-way interaction of group and visual articulation [d] (β = 188.013, t = 2.399, p = .01) showing that Indian English participants produced [d] with a smaller F2-F3 difference (greater retroflexion) when compared to the American English participants.

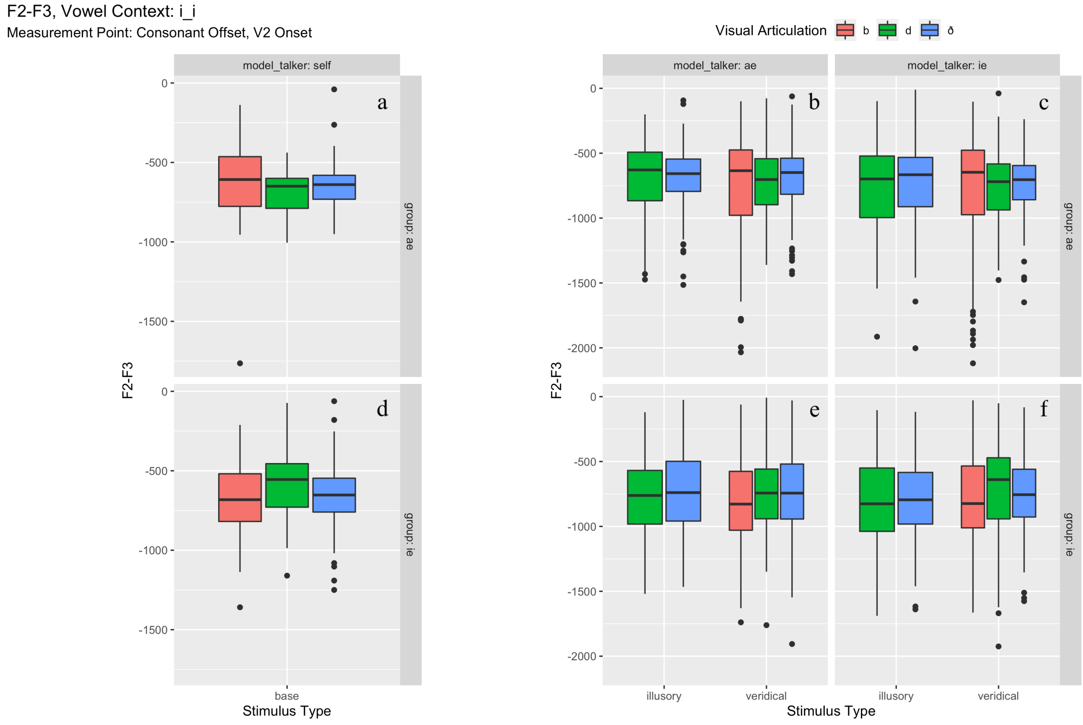

Figure 3.6: Box plot of F2-F3 by stimulus type (baseline, illusory, veridical) for the i_i vowel context measured at consonant offset and V2 onset. The top row shows American English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English). The bottom row shows Indian English participant F2-F3 values for each visual articulation (b, d, ð) and model talker (Self, American English, Indian English).

Tukey-adjusted pairwise comparisons tested for possible effects of visual articulation between baseline, veridical and illusory stimulus types. Indian English participants (Table 3.13) were found to show a difference between baseline and illusory [d] shadowed productions, such that illusory productions were more anterior than baseline. No effects were found for American English participants (Table 3.12). Again, widespread effects were not found likely due to the high frequency F2 of /i/.

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /b/ vs. Baseline /d/	77.043	57.9	1.331	0.999
Baseline /b/ vs. Baseline /ð/	22.18	57.9	0.383	1
Within Dialect (AE Model, AE Participant), Vowel Context: i_i, C Offset				
Baseline /d/ vs. Illusory /d/	17.323	47.2	0.367	1
Baseline /d/ vs. Veridical /d/	45.857	47.2	0.971	1
Illusory /d/ vs. Veridical /d/	28.533	33.4	0.854	1
Baseline /ð/ vs. Illusory /ð/	59.393	47.2	1.257	1
Baseline /ð/ vs. Veridical /ð/	70.267	47.2	1.487	0.9994
Illusory /ð/ vs. Veridical /ð/	10.873	33.4	0.325	1
Across Dialect (IE Model Talker, AE Participant), Vowel Context: i_i, C Offset				
Baseline /d/ vs. Illusory /d/	68.677	47.2	1.453	0.9996
Baseline /d/ vs. Veridical /d/	73.877	47.2	1.564	0.9986
Illusory /d/ vs. Veridical /d/	5.2	33.4	0.156	1
Baseline /ð/ vs. Illusory /ð/	92.32	47.2	1.954	0.9701
Baseline /ð/ vs. Veridical /ð/	96.821	47.4	2.044	0.9505
Illusory /ð/ vs. Veridical /ð/	4.501	33.6	0.134	1

Table 3.12: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant offset in /i_i/ context for American English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

Contrast	Estimate	Standard Error	z.ratio	p.value
Baseline /b/ vs. Baseline /d/	-110.97	52.8	-2.1	0.9341
Baseline /b/ vs. Baseline /ð/	-25.16	52.6	-0.478	1
Within Dialect (IE Model, IE Participant), Vowel Context: i_i, C Offset				
Baseline /d/ vs. Illusory /d/	219.714	43.5	5.051	0.0001 *
Baseline /d/ vs. Veridical /d/	123.237	43.4	2.839	0.4437
Illusory /d/ vs. Veridical /d/	-96.476	30.6	-3.15	0.2311
Baseline /ð/ vs. Illusory /ð/	123.1	43.1	2.854	0.432
Baseline /ð/ vs. Veridical /ð/	86.1	43.1	1.996	0.9618
Illusory /ð/ vs. Veridical /ð/	37	30.5	0.542	1
Across Dialect (AE Model Talker, IE Participant), Vowel Context: i_i, C Offset				
Baseline /d/ vs. Illusory /d/	_189.21_	_43.4_	_4.359_	_0.0036 *_
Baseline /d/ vs. Veridical /d/	158.599	43.4	3.654	0.0541
Illusory /d/ vs. Veridical /d/	-30.611	30.5	-1.004	1
Baseline /ð/ vs. Illusory /ð/	64.272	43.1	1.49	0.9994
Baseline /ð/ vs. Veridical /ð/	69.572	43.1	1.613	0.9977
Illusory /ð/ vs. Veridical /ð/	5.3	30.5	0.174	1

Table 3.13: Pairwise comparisons for effects of stimulus type on F2-F3 at consonant offset in /i_i/ context for Indian English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

3.3.5 Results for labials, /a_a/ context

To assess differences between /v/-like and /w/-like articulations, F2n, the normalized F2 measure (§ 3.4), was taken at the midpoint of labial consonant articulations. The results for this measure in the /a_a/ context are given in Figure 3.7 below. The model revealed main effects of model talker (β = -0.09727, t = -2.073, p = 0.0385) and visual articulation (β = -0.24725, t = -5.380, p < .0001). Visual articulation also entered into two-way interactions: (i) with group (β = 0.14298, t = 2.289 p = 0.0223) such that, compared to American English participants, Indian English participants produced [w]/[ʋ] with a higher F2n closer to [v]/[ʋ], (ii) model talker (β = 0.19354, t = 2.947, p = 0.0033) such that participants shadowed [w]/[ʋ] productions with a higher F2n for the Indian English model talker than for the American English talker, and (iii) phase (β = 0.18413, t = 2.002 p = 0.0456) such that illusory [w]/[ʋ] productions were produced with a higher F2n.

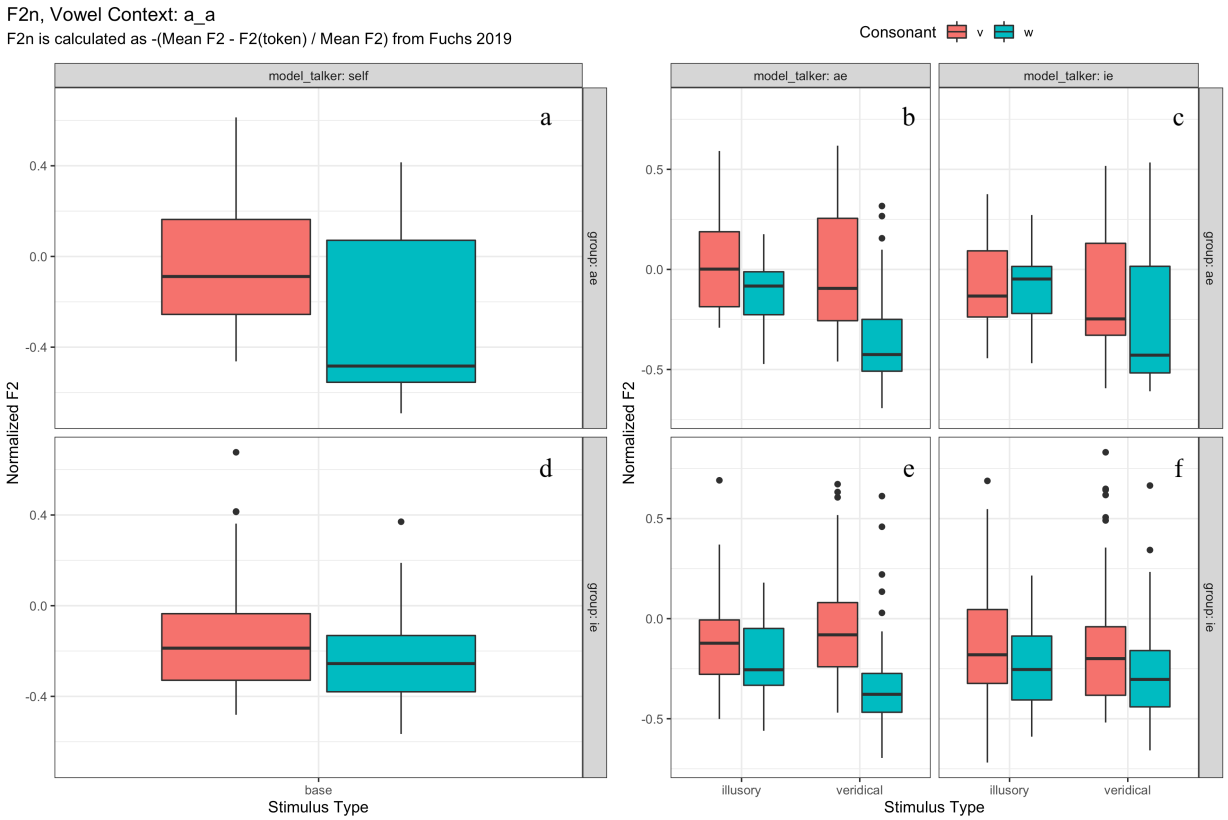

Figure 3.7: Box plot of Normalized F2n by stimulus type (baseline, illusory, veridical) in the a_a vowel condition. A lower F2n value denotes a production with greater velar constriction and lip rounding. The top row shows American English participant F2n values for each visual articulation (v, w) and model talker (Self, American English, Indian English). The bottom row shows Indian English participant F2n values for each visual articulation (v, w) and model talker (Self, American English, Indian English).

Further differences were probed with Tukey-adjusted pairwise comparisons testing for possible effects of visual articulation on participants' labial productions. The outputs of the baseline, veridical, and illusory comparisons can be found in Table 3.14 for American English participants and Table 3.15 for Indian English participants. For American English participants, it was predicted that they would enter the task with separate categories for /v/ and /w/ and that they would maintain these category differences when shadowing a model talker within their dialect. This category preservation, as measured by F2n, is clearly maintained in baseline and veridical trials within dialect (starred rows in Table 3.14) but, unexpectedly, there is no such category maintenance in illusory conditions within dialect suggesting a difference between illusory shadowing and veridical shadowing. Also, as predicted for across dialect conditions, American English participants shifted their productions, with veridical stimuli, in a way that fails to maintain the F2n differences between [v] and [w] found for their baseline productions (compare final rows of within/across sections of Table 3.14). As with the within dialect conditions, there is no measurable F2n difference with illusory stimuli across dialects, however, given the lack of a difference for these stimuli in the within dialect conditions, it is unclear whether this represents a shift to model talker targets akin to what American English participants are doing with veridical stimuli or whether it is again a case of illusory stimuli not engaging shadowing in an equal fashion.

Moving to Table 3.15, Indian English participants showed the same predicted pattern of no shift within dialect but a shift across dialects and did so in the opposite direction of the American English participants (Table 3.14). As seen in Table 3.15, Indian English participants showed no difference in F2n in their [v]/[ʋ] and [w]/[ʋ] baseline productions as well as no difference in F2n when shadowing the Indian English model talker. This was expected given the merger of /v/ and /w/ in Indian English. However, when shadowing the American English model talker in veridical conditions, Indian English participants show a difference in F2n for their [v]/[ʋ] and [w]/[ʋ] productions. Like with the American English participants, this difference is not seen in illusory conditions, again suggesting a difference between veridical and illusory shadowing.

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /v/ vs. Baseline /w/	*0.24725*	*0.046*	*5.38*	*<.0001 **
Within Dialect (AE Model, AE Participant), Vowel Context: a_a				
Baseline /v/ vs. Illusory /v/	-0.0553	0.0492	-1.123	0.9999
Baseline /v/ vs. Veridical /v/	-0.01571	0.0464	-0.338	1
Illusory /v/ vs. Veridical /v/	0.03959	0.0497	0.797	1
Baseline /w/ vs. Illusory /w/	0.06513	0.046	1.417	0.9976
Baseline /w/ vs. Veridical /w/	-0.1586	0.0775	-2.046	0.8833
Illusory /w/ vs. Veridical /w/	0.22373	0.0775	2.887	0.2978
Illusory /v/ vs. Illusory /w/	0.22373	0.0794	1.813	0.9603
Veridical /v/ vs. Veridical /w/	*0.32809*	*0.0464*	*7.064*	*<.0001 **
Across Dialect (IE Model Talker, AE Participant), Vowel Context a_a				
Baseline /v/ vs. Illusory /v/	-0.00449	0.0486	-0.092	1
Baseline /v/ vs. Veridical /v/	0.08156	0.0464	1.756	0.9711
Illusory /v/ vs. Veridical /v/	0.08605	0.0491	1.753	0.9717
Baseline /w/ vs. Illusory /w/	-0.03115	0.046	-0.678	1
Baseline /w/ vs. Veridical /w/	-0.22632	0.0667	-3.392	0.0807
Illusory /w/ vs. Veridical /w/	0.19517	0.0667	2.925	0.2743
Illusory /v/ vs. Illusory /w/	*0.02543*	*0.0683*	*0.372*	*1*
Veridical /v/ vs. Veridical /w/	*-0.06062*	*0.0671*	*-0.904*	*1*

Table 3.14: Pairwise comparisons for effects of stimulus type on F2n in /a_a/ context for American English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /v/ vs. Baseline /w/	*0.14105*	*0.0762*	*1.85*	*0.9517*
Within Dialect (IE Model, IE Participant), Vowel Context: a_a				
Baseline /v/ vs. Illusory /v/	-0.03281	0.0452	-0.727	1
Baseline /v/ vs. Veridical /v/	-0.01423	0.0421	-0.338	1
Illusory /v/ vs. Veridical /v/	0.01858	0.045	0.413	1
Baseline /w/ vs. Illusory /w/	0.02019	0.0421	0.479	1
Baseline /w/ vs. Veridical /w/	-0.02598	0.0526	-0.494	1
Illusory /w/ vs. Veridical /w/	0.04617	0.0524	0.881	1
Illusory /v/ vs. Illusory /w/	*0.1111*	*0.0546*	*2.036*	*0.8882*
Veridical /v/ vs. Veridical /w/	*0.13869*	*0.042*	*3.306*	*0.104*
Across Dialect (AE Model Talker, IE Participant), Vowel Context: a_a				
Baseline /v/ vs. Illusory /v/	-0.03936	0.0442	-0.89	1
Baseline /v/ vs. Veridical /v/	-0.10139	0.0425	-2.386	0.6725
Illusory /v/ vs. Veridical /v/	0.06204	0.0444	-1.396	0.998
Baseline /w/ vs. Illusory /w/	0.08095	0.0421	1.921	0.9317
Baseline /w/ vs. Veridical /w/	-0.06531	0.0593	-1.102	0.999
Illusory /w/ vs. Veridical /w/	0.14626	0.0592	2.471	0.6068
Illusory /v/ vs. Illusory /w/	0.07832	0.0606	1.293	0.9993
Veridical /v/ vs. Veridical /w/	*0.28661*	*0.0423*	*6.772*	*<.0001 **

Table 3.15: Pairwise comparisons for effects of stimulus type on F2n in /a_a/ context for Indian English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

3.3.6 Results for labials, /i_i/ context

Figure 3.8 gives the results of the F2n model for the /i_i/ vowel context, which showed main effects of model talker (β = -0.07010, t = -2.199, p = 0.02814) and visual articulation (β = -

0.2984, t = -9.04, p < .0001). Overall, [w]/[ʋ] had a lower F2n than [v]/[ʋ] (blue vs. corresponding red boxplots throughout Figure 3.8). However, visual articulation entered into two-way interactions with group (β = 0.02774, t = 6.488 p < .001), model talker (β = .09209, t = 2.053, p = 0.04), and phase (β = 0.1441, t = 2.751 p = 0.0607). Moreover, there was a three-way interaction of phase, group, and visual articulation (β = -0.1413, t = -2.322, p = 0.02048). The source of the three-way interaction is likely that the Indian English participants do not produce F2n differences between [w]/[ʋ] and [v]/[ʋ] in some of the phases whereas American English participants regularly do. These differences emerge more clearly in the pairwise comparisons.

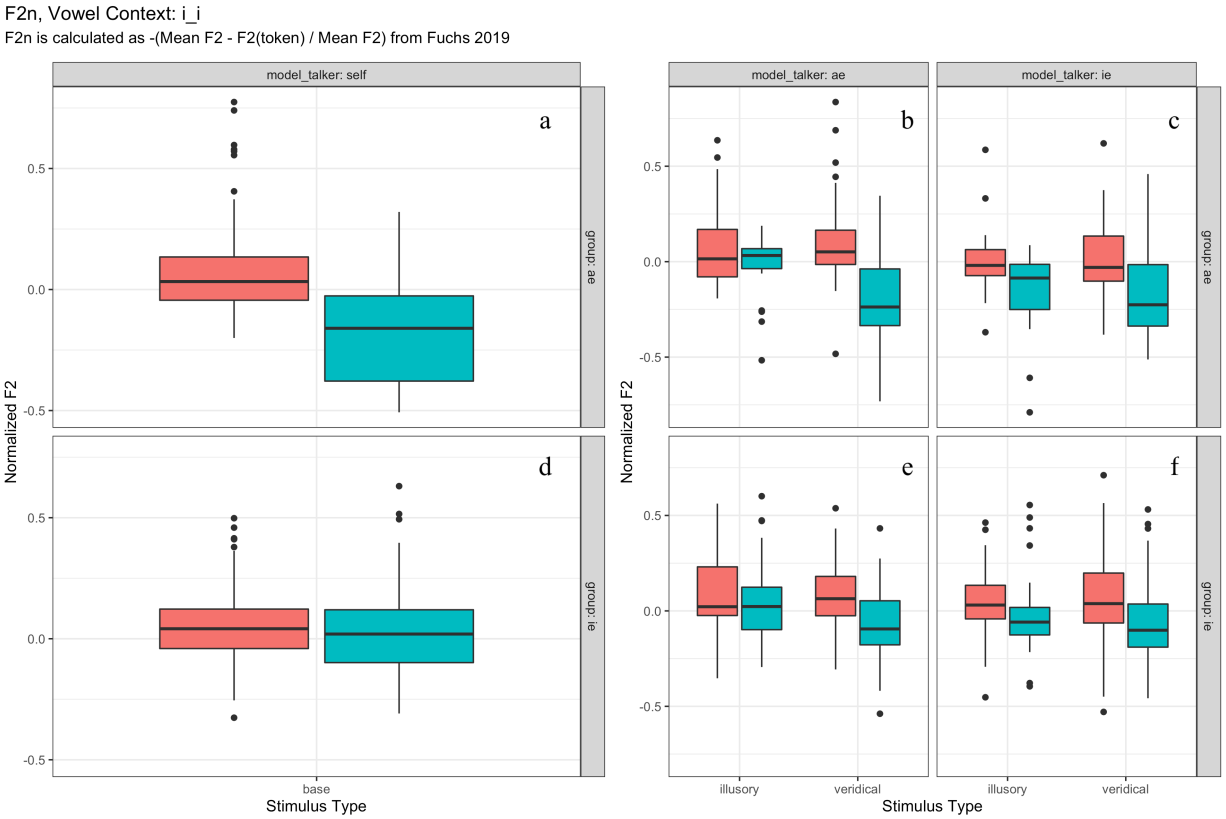

Figure 3.8: Box plot of Normalized F2n by stimulus type (baseline, illusory, veridical) in the i_i vowel condition. The top row shows American English participant F2n values for each visual articulation (v and w) and model talker

(Self, American English, Indian English). The bottom row shows Indian English participant F2n values for each visual articulation (v and w) and model talker (Self, American English, Indian English).

Tukey-adjusted pairwise comparisons were conducted and comparisons for American English (Table 3.16) and Indian English (Table 3.17) are given below. As predicted, American English participants showed a difference in their [v] and [w] productions during the baseline portion of the task. As in the /a_a/ vowel context, these differences in F2n are maintained in the veridical productions of participants when shadowing the within-dialect model talker (Figure 3.8a; veridical). Unlike before, this difference between [v] and [w] in F2n is also found in the illusory productions (Figure 3.8a; illusory) of the American English participants suggesting that they're perceiving some meaningful difference in illusory conditions when they're not failing to McGurk. Unexpectedly, American English participants show a difference in illusory and veridical /w/ (Figure 3.8a; blue boxplots). Across dialects, American English participants also produce F2n differences between [v]and [w] in both veridical and illusory conditions again suggesting that they're perceiving some meaningful and replicable difference in both conditions but at the same time providing evidence that these differences are operating with respect to maintaining underlying category boundaries.

As in the /a_a/ vowel context, Indian English (Table 3.17) participants do not produce a difference between [v]/[ʋ] and [w]/[ʋ] in the baseline task in the /i_i/ context but do produce a difference when shadowing the American English model talker in veridical stimuli. Unexpectedly, Indian English participants also exhibit this difference in (within-dialect) productions for /v/ and /w/ in veridical stimuli when shadowing the Indian English model talker. Additionally, Indian English participants show a difference in F2n between baseline [w]/[ʋ] and across dialect illusory [w] shadowing as well as between across-dialect illusory [w] and veridical [w] productions, suggesting that the veridical [w] stimulus may have a perceptually salient acoustic landmark that facilitates the emergence of shadowed productions that faithfully reflect the category boundaries found in the speech of the American English model talker under veridical conditions.

Contrast	**Estimate**	**Standard Error**	**z.ratio**	**p.value**
Baseline /v/ vs. Baseline /w/	*0.298425*	*0.0314*	*9.504*	*<.0001 **
Within Dialect (AE Model, AE Participant), Vowel Context: i_i				
Baseline /v/ vs. Illusory /v/	0.003699	0.033	0.112	1
Baseline /v/ vs. Veridical /v/	0.015514	0.0317	0.489	1
Illusory /v/ vs. Veridical /v/	0.011814	0.0333	0.355	1
Baseline /w/ vs. Illusory /w/	0.012597	0.0316	0.399	1
Baseline /w/ vs. Veridical /w/	-0.14332	0.0405	-3.536	0.0514
Illusory /w/ vs. Veridical /w/	0.155918	0.0407	3.835	0.0182 *
Illusory /v/ vs. Illusory /w/	*0.151405*	*0.0416*	*3.638*	*0.0365 **
Veridical /v/ vs. Veridical /w/	*0.295509*	*0.0319*	*9.266*	*<.0001 **
Across Dialect (IE Model Talker, AE Participant), Vowel Context: i_i				
Baseline /v/ vs. Illusory /v/	0.074387	0.0362	2.057	0.8784
Baseline /v/ vs. Veridical /v/	0.085613	0.0316	2.713	0.4186
Illusory /v/ vs. Veridical /v/	0.011226	0.0363	0.309	1
Baseline /w/ vs. Illusory /w/	-0.0094	0.0314	-0.299	1
Baseline /w/ vs. Veridical /w/	-0.18555	0.0457	-0.406	1
Illusory /w/ vs. Veridical /w/	0.009159	0.0457	0.2	1
Illusory /v/ vs. Illusory /w/	0.205483	0.0488	4.211	0.0041 *
Veridical /v/ vs. Veridical /w/	0.203416	0.0316	6.447	<.0001 *

Table 3.16: Pairwise comparisons for effects of stimulus type on F2n in /i_i/ vowel context for American English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

Contrast	Estimate	Standard Error	z.ratio	p.value
Baseline /v/ vs. Baseline /w/	*0.020977*	*0.029*	*0.723*	*1*
Within Dialect (IE Model, IE Participant), Vowel Context: i_i				
Baseline /v/ vs. Illusory /v/	0.020319	0.0308	0.66	1
Baseline /v/ vs. Veridical /v/	-0.006506	0.0292	-0.223	1
Illusory /v/ vs. Veridical /v/	-0.026825	0.031	-0.865	1
Baseline /w/ vs. Illusory /w/	0.089051	0.0293	3.04	0.2101
Baseline /w/ vs. Veridical /w/	0.050751	0.338	1.502	0.995
Illusory /w/ vs. Veridical /w/	0.0383	0.339	1.13	0.9999
Illusory /v/ vs. Illusory /w/	0.051409	0.0352	1.461	0.9965
Veridical /v/ vs. Veridical /w/	0.116534	0.0294	3.963	0.0119 *
Across Dialect (AE Model Talker, IE Participant), Vowel Context: i_i				
Baseline /v/ vs. Illusory /v/	-0.024136	0.0301	-0.803	1
Baseline /v/ vs. Veridical /v/	-0.030595	0.0293	-1.044	1
Illusory /v/ vs. Veridical /v/	-0.006459	0.0305	-0.212	1
Baseline /w/ vs. Illusory /w/	0.107779	0.029	3.713	0.0282 *
Baseline /w/ vs. Veridical /w/	-0.23388	0.0327	-0.716	1
Illusory /w/ vs. Veridical /w/	0.131167	0.0326	4.028	0.0087 *
Illusory /v/ vs. Illusory /w/	-0.021724	0.0336	0.647	1
Veridical /v/ vs. Veridical /w/	*0.15935*	*0.0293*	*5.439*	*<.0001* *

Table 3.17: Pairwise comparisons for effects of stimulus type on F2n in i_i vowel context for Indian English participants. Comparisons for which there is an effect are starred. Comparisons that are consistent with predictions are in italics.

3.4 Discussion

The research question at the center of this experiment was whether participants who are tasked with shadowing veridical and illusory multisensory stimuli produced by model talkers from different English dialects show production patterns that mirror their own (the participants') baseline production strategy or those of the model talkers. This was probed via veridical and illusory shadowing through two dependent measures, F2-F3 for /b/, /d/, and /ð/ and F2n for /v/ and /w/, across two vowel conditions. Broadly, participants exhibited baseline productions that mirrored the description of their dialects in the literature. Indian English participants showed a greater degree of retroflexion for coronal productions as measured by F2-F3 and F2n values consistent with a merger of /v/ and /w/ to /ʋ/. American English participants showed more anterior coronal productions as measured by F2-F3 and a difference in F2n values reflective of the phonemic contrast between /v/ and /w/ in American English.

In across dialect shadowing, American English participants were predicted to show a shift for coronals where the difference of F2-F3 diminished, reflecting a more posterior production, as well as a collapse of the difference between labial /v/ and /w/ as cued by F2n reflecting the merger found in Indian English. In across dialect shadowing, Indian English participants were predicted to show a shift in coronal productions where the difference inF2-F3 was enhanced as well as a difference in F2n between /v/ and /w/ reflecting their distributions in American English.

Coronals: In the F2-F3 models, effects of shadowing conditions were largely confined to the /a_a/ vowel context. As noted in §3.4, §3.6.3, and §3.6.4, negligible effects in the /i_i/ context were likely due to the high F2 of /i/ obscuring any task-related differences in F2-F3. Within the /a_a/ vowel context a subset of predictions was borne out, but only for Indian

English participants. These results are presented below in Table 3.18. As noted above, Indian English participants produced baseline [d] tokens that exhibited a more posterior production congruent with [ɖ]. In veridical shadowing, both within and across dialects they kept this production strategy and exhibited no statistical difference between their baseline recordings and veridical shadowed productions. In illusory shadowing, Indian English participants showed a shift towards more anterior articulations that were both statistically different from baseline and veridical shadowed productions. For /ð/, Indian English participants showed statistical differences between baseline and illusory and baseline and veridical shadowed productions but no difference between veridical and illusory shadowed productions. With /d/ it appears as though Indian English participants are using their own production strategies when shadowing and with /ð/ it appears that illusory and veridical shadowed productions are acoustically similar to the exclusion of baseline trends.

Coronal Consonants	**Acoustic Measure**	**Vowel**	**Baseline**	**Veridical Results (Within)**	**Veridical Results (Across)**	**Illusory Results (Within)**	**Illusory Results (Across)**
Indian English /d/	F2-F3	/a_a/	Small F2-F3 difference, posterior production	*No change, posterior production*	No change, posterior production	Larger F2-F3 difference, anterior production	*Larger F2-F3 difference, anterior production*
Indian English /ð/			Small F2-F3 difference, posterior production	Larger F2-F3 difference, anterior production	*Larger F2-F3 difference, anterior production*	Larger F2-F3 difference, anterior production	*Larger F2-F3 difference, anterior production*

Table 3.18: Shadowing results where participants showed a shift relative to baseline in the shadowing of coronal consonants. Italics denote predictions borne out in the results.

One possibility for this difference between the distributions of shadowed /d/ and /ð/ is the status of each of these sounds in Indian English. While /d/ has a phonemic base in Indian English, /ð/ does not and instead is potentially the output of some substitution procedure (e.g. /ð/ → [d], [d̤], etc.). As such, /d/ has a transparent perceptual anchor while /ð/ does not. We can see some evidence of this in Figure 3.3, where baseline [d] productions show less variability than baseline [ð] productions. Another possibility for this difference between /d/ and /ð/ is role of lowering F3 in Indian English. Often questions of social expectation, or illusory perception for that matter, hinge on ambiguity or confusability in the signal. In the case of veridical [ɖ] there is a highly salient, reliable cue (F3 lowering) that allows for little confusability. One possibility is that the perceptual clarity of lowered F3 is what's driving the effects in shadowing to be limited strictly to veridical cases and not illusory trials.

What is most unexpected is that these shifts for Indian English participants occurred in both within and across dialect conditions. For American English participants, there were no shifts in any of the model talker conditions regardless of visual articulation or stimulus type (illusory, veridical; Tables 3.6 and 3.8). That the American English participants produced no shifts for the Indian English model talker runs counter to predictions, but is perhaps especially unexpected given that Indian English participants showed shifts in both model talker conditions (Tables 3.7 and 3.9). While a majority of all participants, and nearly every Indian English participant, reported thinking that the experiment was about accents and the role that visually seeing the model talker plays in understanding accents, the cultural stakes are different for the American English and Indian English participants. As noted above in § 3.3.2, Indian English participants seemed acutely aware of the fact that Indian English, or Indian accents, are sociolinguistically marked when compared to other Englishes. Not only this but, anecdotally, different Indian English participants showed metalinguistic awareness that some of their productions in the practice trials, despite being cued with different orthography, were essentially the same sounds. Given the relationship between Indian English and American English, there is an argument to be made that Indian English participants were simply more aware of the kinds of phonological substitutions they make on a daily basis and given their lived reality and language usage. There is also an argument to make via Kutlu (2020) that the listening environment that Indian English participants exist in, particularly given their immigrant status in the US and Canada, provides them with a wider array of sociophonetic experience when compared to a linguistically more homogenous experience like the one American English participants receive. I return to these ideas in Chapter 4 in more detail about why different patterns might be emerging for /d/ and /ð/ in Indian English shadowed productions.

Labials: In the F2n models there were effects of task conditions in both vowel contexts as shown in Table 3.19 below. The /a_a/ context provided somewhat tidier results given the predictions presented in §3.5. In the /a_a/ context, American English participants produced a difference in F2n between [v] and [w] in baseline recordings that was maintained when shadowing the American English model talker with veridical stimuli but collapsed elsewhere, most crucially—and as expected—when shadowing veridical and illusory stimuli of the Indian English model talker (Table 3.14 and Table 3.19). Indian English participants also showed a shift when shadowing across dialect. Crucially, as expected, Indian English participants did not differentiate [v] and [w] in F2n in baseline recordings, and only produced a (F2n) /v/-/w/ difference when shadowing the American English model talker in veridical conditions (Table

3.15 and Table 3.19). The green highlighted cells in Table 3.19 highlight these within/across dialect shifts within the veridical trials for both participant groups. Taken together these patterns show compelling evidence of across dialect shift in the veridical stimuli condition. While both groups also showed some predicted outcomes in the illusory stimuli condition, there is a glaring lack of /v/ - /w/ contrast preservation (American English; illusory within) and facilitation (Indian English; illusory across). The veridical results serve as support to what is already known about shadowing, namely that participants move towards the targets of model talkers. However, it leaves open questions about the lack of shift in illusory conditions when shadowing across dialect boundaries.

Labial Consonants	Acoustic Measure	Vowel	Baseline	Veridical Results (Within)	Veridical Results (Across)	Illusory Results (Within)	Illusory Results (Across)
Indian English	F2n	/a_a/	No contrast between /v/ and /w/	*No contrast between /v/ and /w/*	*Contrast between /v/ and /w/*	*No contrast between /v/ and /w/*	No contrast between /v/ and /w/
American English			Contrast between /v/ and /w/	*Contrast between /v/ and /w/.*	*No contrast between /v/ and /w/.*	No contrast between /v/ and /w/.	*No contrast between /v/ and /w/.*
Indian English	F2n	/i_i/	No contrast between /v/ and /w/	Contrast between /v/ and /w/	*Contrast between /v/ and /w/*	*No contrast between /v/ and /w/*	No contrast between /v/ and /w/
American English			Contrast between /v/ and /w/	*Contrast between /v/ and /w/*	Contrast between /v/ and /w/	*Contrast between /v/ and /w/*	Contrast between /v/ and /w/

Table 3.19: Shadowing results where participants showed a shift, relative to baseline, in the shadowing of labial consonants. Italics denote conditions where predictions were borne out in the results. Green shading highlights the results of Indian English and American English veridical shadowing where each participant group shows complementary patterns to one another.

For the F2n models within the /i_i/ context, the story is more complicated but provides useful insight into how illusory and veridical stimuli differ from one another. Like in the /a_a/ vowel condition, American English participants produced a difference in F2n between [v] and [w] in the baseline condition, however this time it was maintained, across the board under all conditions (last row of Table 3.19). Of note is that American English participants unexpectedly, maintained this difference between [v] and [w] when shadowing the Indian English model talker in both the illusory and veridical conditions. This differs from the /a_a/ condition where it appeared as though American English participants were utilizing model talker acoustic patterns as their object of veridical imitation. This could again be reflective of a tradeoff between perceptual confusability and signal clarity in the stimuli. In the case of the /a_a/ context, as was the case in Chapter 2, there are gestural targets and vowel contexts that yield not only clear production landmarks but also wide jaw apertures that facilitate the tracking of gestural movements. In the /i_i/ context, this clarity is reduced, facilitating not only an illusion but one that also shows evidence of top down expectational perceptual learning (e.g. in the case of within dialect illusory shadowing of /v/ and /w/ by American English participants).

Indian English participants' results showed a similar pattern in the /i_i/ context. Again, like in the /a_a/ vowel condition, Indian English participants did not produce a difference in F2n between baseline [v]/[ʋ] and [w]/[ʋ] but, as predicted, that difference emerged in the veridical cross dialect condition when shadowing the American English model talker (Table 3.19). However, Indian English participants extended this emergent contrast in veridical shadows of the Indian English model talker (who was predicted to have a merged F2n). In addition to these patterns, Indian English participants also showed variation in their [w] productions (Table 3.17), which exhibited differences between baseline and illusory, and illusory and veridical, conditions. This provides further support for Fuchs (2019) who argues that F2 is the primary cue being utilized by Indian English speakers, but also these results are akin to the pattern found for F2-F3 /d/ for Indian English participants where the baseline and veridical productions align with one another to the exclusion of the illusory.

One broad takeaway from this experiment is that illusory and veridical shadowing do not appear to work in the same way. This is seen in the different patterns of shadowing results for /d/ and /ð/ in the /a_a/ context by Indian English participants as well as the shadowed /v/ and /w/ productions in both contexts by American English participants. For Indian English participants, when they produce a significant shift between baseline and shadowing, /d/ and /ð/ productions appear to behave differently: for /d/, veridical shadowed and baseline, but not illusory shadowed, productions are the same but for /ð/, veridical and illusory shadowing (for both within and across conditions) differ from baseline productions (e.g. /ð/). In the case of /d/ it seems somewhat straightforward to say that illusory stimuli simply do not behave like veridical stimuli and that participants are likely using something akin to their baseline productions while veridically shadowing (Mitterer & Mussler, 2013) . While it is tempting to say that the shadowing seen in /ð/ is the product of sociophonetic experience, given the broad deviation from baseline production patterns, it remains to be seen. As noted above, this is returned to in more detail in Chapter 4.

In the case of American English participants shadowing /v/ and /w/, the finding that veridical and illusory stimuli elicit different response patterns again emerges. In the /a_a/ context, predicted patterns of (within dialect) listener experience affecting productions were only seen in veridical shadowing. This is consistent with different degrees of perceptual clarity between illusory and veridical percepts. However, in the /i_i/ context, there is evidence that illusory shadowing and veridical shadowing are the same, but differ from baseline. As noted above, the /a_a/ context presents the more compelling case given the clear collapse/maintenance of a contrast across different model talkers. Whether the results for /i_i/ are evidence of perceptual similarity across trial types (veridical/illusory) or an artifact of the perceptual confusability/clarity of the stimuli remains to be seen. Teasing this apart through further experimentation could yield results that could speak to whether sociophonetic expectations are fundamentally linked to acoustics (hence the veridical but not illusory shifts in the /a_a/ context) or whether expectational processes can persist despite a lack of acoustic reinforcement. In Chapter 4 I will present the main takeaways from both of experiments and situate my findings more broadly in the literature.

Chapter 4 Discussion

The primary focus of this dissertation was understanding whether socially indexed listener expectations are dependent on acoustic signals. To explore this question I used illusions as a strategic tool to probe whether English-speaking listeners, when exposed to the systematic speech patterns of a speaker of another variety of English, were more likely to arrive at illusory percepts that were faithful to their own phonology or that reflected the phonology of the other variety. Given that listeners are generally adept perceptual learners when exposed to novel speech patterns (Norris, McQueen & Cutler 2003, Shockley et al. 2004, Kraljic & Samuel 2006, Nielsen 2011, Trude & Brown-Schmidt 2012) illusions were crucial for probing the question of acoustic dependency because illusory stimuli provide the *wrong* type of acoustic information to listeners. Rather, listeners must rely on the visual signal, which contains linguistically and socially relevant gestural information, in tandem with their phonological knowledge to arrive at a meaningful percept.

Two experiments were conducted to probe this question with speakers from two varieties of English, Indian English and American English. Experiment 1 investigated the categorization strategies of Indian English and American English listeners when they were presented with illusory stimuli before and after gaining experience with the veridical speech patterns of Indian English and American English model talkers. Experiment 2 analyzed the production strategies of participants between their baseline productions of target stimuli and their productions when asked to shadow veridical (audio-visual congruent) and illusory (audiovisual incongruent) stimuli from the model talkers. Taken together, results from these two experiments yield explicit and implicit measures of consonantal identity – providing evidence of the degree to which listener experience and social expectation can persist despite a lack of acoustic reinforcement. Section §4.1 discusses specific findings for each group and 4.2 discusses general findings across the two experiments.

4.1 Group Results

4.1.1 American English Results

When interpreting the speech of the Indian English model talker, American English participants came to both experimental tasks with a grammar that required them to condense a larger set of contrasts into a smaller one either by means of 1) substitution of one sound for another (e.g. /ð/ → [d]) or 2) the merger of two existing contrasting units into a single non-contrastive unit (e.g. /v, w/ → [ʋ]). American English participants appeared to be able to achieve this end as a result of experience with the model talker for a subset of labial conditions in both categorization and shadowing. Recall that, for all illusory stimuli, the acoustic signal was always [b] which lacks the acoustic cues associated with productions of [d], [ḍ̤], [ð], [v], [w], and [ʋ]. In categorization (Experiment 1), American English participants only showed an across-dialect effect of the intervention video for labials in the /a_a/ context, where they were more likely to categorize the Indian English model talker's /w/ (visual [ʋ]) productions as /v/ post-intervention. Given the near merger of /v/ and /w/ in Indian English, it seems as though American English participants were utilizing a form of perpetual learning rooted in their experience with the Indian English model talker. This was the only instance of American

English participants shifting their categorization strategy as a result of their experience across dialects.

In shadowing (Experiment 2), American English participants again showed an across-dialect effect of the intervention video for labials. Although these participants preserved the F2n contrast between /v/ and /w/ in their baseline and veridical shadowed productions for the American English model talker, they did not when shadowing veridical stimuli from the Indian English model talker, reflecting the merger of these sounds in Indian English (Table 3.19). This is the result one would expect in a traditional shadowing framework, where participants show the ability to actively imitate the speech of the model talker that they are shadowing. Although the across-dialect result also appeared to extend to the illusory condition, where American English participants shadowed the illusory labial stimuli in the /a_a/ context from the Indian English model talker with merged F2n values, the lack of contrast preservation in the corresponding within-dialect illusory condition for the American English model talker (Table 3.19) leaves doubt as to whether the merger with the Indian English model talker is reflective of speaker-specific learning.

This picture is further complicated when considering American English participants' within-dialect *categorization* of these stimuli, where American English participants experienced more McGurk effects for labials after intervention in the /a_a/ context *and* categorized these illusory percepts as [w]. Given American English participants' experience with the American English model talker, there should be nothing for the burst of acoustic [b] to map onto when listeners are presented with a visual articulation for [w] produced by the American English model talker. Because the McGurk effect hinges on the possibility that listeners misinterpret the transient burst of [b] as being a part of the visually articulated speech gesture, and because

American English /w/, unlike Indian English /ʋ/, lacks acoustic information consistent with that misinterpretation in the veridical world, American English participants were not expected to McGurk with the American English model talker.

As I argue in §2.4.2 I believe this is a case of American English participants lip-reading and using the clearest signal they have to complete the task. In this formulation, listeners, upon 'hearing' an insufficiently clear percept, pivot to identifying it with the clearest signal at their disposal (the visual signal). Arguably, this is different from what these same participants are doing in the Indian English model talker case where there is phonological and sociolinguistic support for categorizing Indian English [ʋ] as /v/ after intervention. These results are congruent with both associative theories of multisensory perception where perception is argued to largely depend on associative experience between the auditory and visual streams (Shams, 2011, Magnotti & Beauchamp 2017) and supramodal theories of multisensory perception that argue for a modality neutral perceptual apparatus that uses shared information across modalities to arrive at percepts (Fowler 2004, Rosenblum et al., 2016).

A key theoretical difference at stake is whether the results for American English categorization are the product of one or two perceptual channels, which this experiment does not tease out. Under a supramodal interpretation of the results, American English participants perceived different consonants (i.e. /w/ for the American English model talker, /v/ for the Indian English model talker) because the articulations of [w] and [ʋ] contained different shared information across their productions. Put another way, a single unified perceptual percept reflected the inherent differences in the visual stimuli. Under an associative interpretation, the different results *could* stem from different perceptual information channels. In the case of American English [w], if the perceptual content wasn't clear enough, participants *could* reanalyze the stimulus using the visual signal as a backup to audition – a late arrival at a percept. In the case of Indian English visual [ʋ], participants had sufficient associations between the acoustics and visual articulations of [ʋ] to facilitate a straightforward percept. Taken together, it appears as though participants, under the right conditions, namely categorization, were able to rely on their sociolinguistic experience with the natural speech of the model talkers to facilitate speaker specific illusory perception without a congruent or strong reinforcement of the acoustic signal.

Despite cross dialect intervention effects being limited to labial consonants in the /a_a/ context, American English participants also responded differently to the stimuli produced by the Indian English and American English talkers in ways that appear to be sensitive to the broader phonological patterns of these varieties. In categorization (Experiment 1), American English participants were more likely to report perceiving /ð/ when viewing the visual articulation of [ð] from the American English model talker and /d/ (Figure 2.5a & 2.5b) when they were presented with visual articulations for [d̤] from the Indian English model talker in the /a_a/ condition. Similarly, in the /i_i/ condition with labial consonants, American English participants were more likely to perceive /v/ when presented with visual articulations for [w] by the Indian English model speaker. Throughout Experiment 1 (i.e., even pre-intervention), American English participants' responses are consistent with an awareness of how certain consonants of Indian English and American English might sound and, in specific cases described above, this awareness is further enhanced (as shown by categorization shifts) by the speaker-specific experience they receive from the intervention video. This mirrors other work in the veridical literature that suggests that experience with a given speaker can contribute to adjustments in listeners' perceptual decisions (Kraljic & Samuel 2006, Dahan, Drucker &

Scarborough 2008, Kraljic, Brennan & Samuel 2008, Samuel & Kraljic 2009, Coetzee et al. 2022).

Taken wholistically, there is evidence that, as American English participants gained experience with the Indian English model talker's speech patterns across the course of the experiment, they utilized what they had learned to facilitate shifts in illusory categorization and veridical shadowing. For example, the visual articulation for /w/—here, [ʋ]—paired with the experience of perceiving a speaker for which the sounds /v/ and /w/ are merged in their natural speech patterns, is sufficient to facilitate a McGurk effect reflective of that perceptual experience. Importantly though, this effect appears to be limited to linguistic experience where /w/ can have confusable characteristics with /v/ which it does in Indian English [ʋ]. Despite the potential of American English participants using different listening modes across model talkers in conditions with visual articulations for /w/, when taken together, the fact that they showed different patterns of categorization suggests that they are sensitive to the phonetics of [v], [w], and [ʋ] in the natural productions of the model speakers and are using this information when categorizing illusory stimuli despite the lack of acoustic reinforcement. While this result doesn't shed light on whether this effect is primarily driven by reliably stable and identifiable gestural landmarks in the visual signal or experiential knowledge about real world variability in perception (i.e. the fact that American English listeners sometimes perceive Indian English 'wave' as [weɪv] and sometimes as [veɪv]), this result suggests that linguistic experience with the real world phonological patterns of a given speaker contributes to illusory perception and can serve to facilitate sociophonetically grounded categorization strategies under otherwise adverse listening conditions.

4.1.2 Indian English Results

Unlike American English participants, Indian English participants came to both experimental tasks with a grammar that required them to expand a smaller set of contrasts to a larger set of contrasts by means of splitting single non-contrastive units (e.g. /d/, /ʋ/) into sets of contrasts (e.g /d/, /ð/, and /v/, /w/ respectively) when perceiving the American English model talker. Within the categorization task, Indian English participants did not do this. While intervention did facilitate more illusions generally, Indian English participants showed no across dialect categorization shifts after intervention.

In shadowing, Indian English participants showed across dialect production differences in F2-F3 for the coronals between baseline, veridical and illusory trials for the American English model talker. However, all of the across dialect differences seen for the American English model talker were also seen within dialect for the Indian English model talker. Indian English participants appear to be using two distinct production strategies in the coronal /a_a/ context which are schematized in Figure 4.1. With /d/, Indian English participants show shifts in F2-F3 such that baseline and illusory and illusory and veridical productions were statistically different from one another (Table 3.18, Figure 4.1a). This suggests that baseline and veridical shadowed productions showed more similarity with one another to the exclusion of shadowed productions from illusory stimuli. This difference was maintained both at the consonant onset and offset for both model talkers. For /ð/, baseline productions differed in F2-F3 from both veridical and illusory shadowed productions with no difference between illusory and veridical shadows (Table 3.18, Figure 4.1b). Put another way, shadowed productions were more similar to one another to the exclusion of baseline tokens.

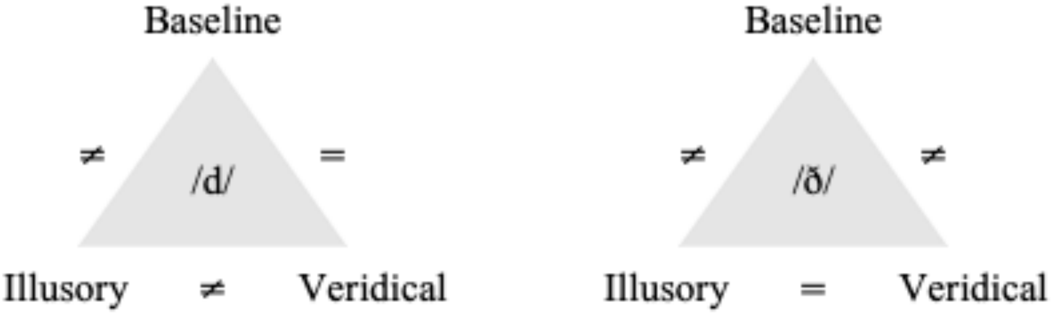

Figure 4.1: Schematization of Indian English production strategies in coronal /a_a/ contexts for both model talkers.

What is unclear is why these differing patterns emerge for the Indian English participants. In the across dialect /d/ veridical conditions, it is curious that Indian English participants don't produce more anterior productions in the veridical trials given the F2-F3 of the American English model talker (Figure 3.2). Thinking back to §1.2, Mitterer and colleagues (2008, 2013) note that shadowers often use baseline productions unless presented with a socially salient phonological variant. In this case, the production patterns of the Indian English model talker for /d/ are presumably no different from those of the Indian English participants and so it is unsurprising that baseline and veridical trials pattern together. However, the lack of a shift when shadowing the American English model talker may suggest the American English dialect is not socially salient, or perhaps socially relevant, enough for Indian English speakers to shift their production strategies from baseline. As noted in §1.4, while Indian English may have had Received Pronunciation (RP) phonological targets in early language contact, it is now a phonologized form and is increasingly showing localized geographic variation (Gargesh 2008, Sailaja 2012, Wiltshire 2020). That within dialect variation is likely where Indian English participants are doing most of their sociolinguistic meaning making. As such, within dialect variation may hold a more salient place in the sociolinguistic awareness of Indian English participants than across dialect variation. This is different from American English sociolinguistic pressures where Indian English accents hold a marked status as deviant from an American English norm. While American English listeners may notice the *type* of /d/ produced by Indian English speakers, viewing their own American English dialect as privileged within a Global English context, Indian English talkers may perceive the /d/ productions of the American English model talker as being another /d/-like token among the many variations across the array of Global Englishes.

This pattern differs notably from /ð/ shadows where illusory and veridical shadowed productions are more similar to one another in F2-F3 to the exclusion of baseline. A straightforward interpretation of the result is that there is a perceptual equivalency between illusory and veridical percepts despite what was seen with the patterning in /d/. While this is tempting, I am hesitant to assert this given that Indian English participants showed more anterior shadowed productions across both model talkers and both shadowing contexts (illusory, veridical). Indian English participants also produced within-dialect categorization shifts in a direction that suggest Indian English participants expected that the Indian English model talker had American English phonology. As noted in § 2.2.1, the natural productions of [d̤] by the Indian English model talker have visible tongue/tooth contact and are susceptible to being perceived as /ð/ when paired with the acoustics for [b]. The fact that Indian English participants, who arguably do not have a phonemic category for /ð/, lead with a categorization strategy that prioritizes /ð/ and then shifts to /d/ suggests that are sensitive to the talker specific patterns they were exposed to during the intervention phase.

Why Indian English participants came to task with matching expectations for both model talkers warrants further investigation but one speculative and testable answer, mentioned in the discussion to Chapter 2, is that Indian English participants, despite performing the task in India, may have been in an American English listening mode given how participants accessed the experiment (Amazon Mechanical Turk) and the University of Michigan branded consent paper work that they were required to complete before beginning the experiment proper. In this way, participants may have thought that the content of the experiment was about English as spoken in the United States and attuned their listening ‘appropriately’ before the task started. Given that all stimuli in the pre-intervention block were illusory (except for veridical check trials of [b]), Indian English participants wouldn’t know that the Indian English model talker spoke Indian English as opposed to being an Indian American model talker speaking American English. If this was the case, one would expect to see exactly what Indian English participants did in Experiment 1, which was shift their categorization strategies after intervention to Indian English phonological patterning. This pre-intervention pattern, like the one seen above with American English participants’ responses, is consistent with an awareness of how certain consonants of American English sound and this awareness is further enhanced (as shown by categorization shifts) by the speaker-specific experience they receive from the intervention video. There is also the possibility that Indian English participants, all of whom reported speaking another language other than English in the home as a child, are performing as though this is an L2 perceptual task. That is, while they may have perceptual acuity for within and across category differences in their various L1’s, this ability within their L2 may have been impacted by their L1 (Flege & Bohn, 2021). For example, if it is the case that the Indian English participants in this study had no /ð/ category in their L1, it should come as little surprise that their strategy for shadowing /ð/ differs so dramatically from their strategy for /d/

which a subset of participants presumably have in the inventory of their L1 (e.g. Hindi, Malayalam).

4.2 General Findings

Generally, the findings demonstrate that the speaker-specific expectations that listeners build via their language experience *can* play a role in determining the content of the illusory percepts they experience. This indicates that listener experience plays a role in percept construction even when the acoustic signal fails to strongly reinforce the visual signals that participants are exposed to. It is worth noting however that, in Experiment 1, this effect is constrained to the broad measures provided by categorization, which isn't a useful measure of within category perceptual sensitivity. To gain a better understanding as to whether this same kind of experience gave rise to within category variation for illusory stimuli, shadowing was tested in Experiment 2. In the shadowing results, though, listeners' experienced-based shifts (relative to their baseline productions) were largely confined to veridical conditions. The veridical results suggest that listener experience is still at play in shadowing, as is consistent with the broader perceptual learning and accommodation literature (Shockley et al. 2004, Mitterer & Ernestus 2008, Honorof et al. 2011, Nielsen 2011, Mitterer & Müssler 2013, Kwon 2019). However, for illusory trials the perceptual conditions that illusory stimuli provide may be too adverse due to their incongruent nature for listeners to overcome or for listener experience to prove useful.

These takeaways are not without their caveats though. In both experiments, post-intervention trials obviously always followed pre-intervention trials potentially introducing a confound of ordering. In some contexts, such as American English participants categorizing

American English [w] and Indian English [v]/[ʋ], there was evidence of a dissociation that lent support to the fact that listeners appeared to be using sociophonetic experience in their categorization strategy. In other cases where a dissociation between stimuli wasn't found (e.g. illusory shadowing of [v]/[w]/[ʋ] for both groups) it is unclear whether these one sided effects are the product of sociophonetic experience or just familiarity with the stimuli.

Also, given that Gentilucci & Cattaneo (2005) were able to find reliable effects of one modality (auditory or visual) affecting the other modality in participant shadowing, it begs the question why effects in this study were limited to only veridical conditions. One possibility is that the small number of participants in the shadowing experiment (n = 23) didn't yield enough 'good' shadowers (Gentilucci & Cattaneo used 65 participants in their work). Because shadowing is notoriously sensitive to individual variation (Babel 2012), it can be the case that with enough "poor" shadowers one washes out the effect found with those who shadow effectively. However, based on an informal look at individual performance in Experiment 2 shadowing, it appears that many while many participants were effective shadowers in veridical conditions, they only extended their across or within dialect patterns to at most one of the illusory conditions. As noted above, without a dissociation between strategies for each of the model speakers, it's unclear whether shadowing in the illusory trials was the product of perceptual adaptation or not.

The broader categorization findings that sociolinguistic expectations can shift categorization strategies even in illusory contexts appear to be congruent with both gesturalist and acoustic theories of speech and multisensory perception. In particular, both Direct Realist and acoustic speech perception theories seem capable of handling the categorization findings. For Direct Realism, as Fowler (1996) notes, when a listener can be tricked into believing that there is a single distal source, they will experience a McGurk effect. In this case, the distal source appears to include speaker specific gestural information that influences the categorization patterns of listeners. As for acoustic theories, and with them exemplar models of speech perception, these results suggest a need for incorporating visual processing of some sort into the exemplar frameworks. Because McGurk effects appear to be driven by a visual signal that crucially eclipses the acoustic signal, an acoustic exemplar framework for speech perception that is solely acoustic is woefully incomplete.

In addition to these larger takeaways, both in the categorization and shadowing experiments, vowel condition appeared to play a significant role in the degree to which participants experienced illusions, in ways again consistent with the literature (Green & Kuhl 1991, Burnham 1998, Shigeno 2002, Burham & Dodd 2017). While not designed as a fixed effect, participants in both groups appeared to experience more illusions in /a_a/ vowel conditions as opposed to /i_i/ conditions. As noted throughout this dissertation, the trend for increased illusory perception within the /a_a/ vowel condition is likely due to more gestural information from the greater oral aperture associated with [a] as compared to [i]. Labial conditions with /w/ and /v/ also appeared to elicit larger effects across the two experiments likely due to either the clarity of the gestures given their oral articulation or the sociolinguistic markedness of the merger of /v/ and /w/ as compared to the substitution of [d] for /ð/. While the fortition of /ð/ → [d] is fairly common across the World's Englishes (Zhao 2010), the merger of /v/ and /w/ is not and both American and Indian English speakers appear to be aware of this either due to the perception of lexical errors (e.g. [veɪvz] or [weɪvz] for 'waves') on the part of American English listeners or the experience of being corrected or asked to repeat what was said on the part of Indian English speakers. As noted above, one novel contribution of this work is showing that visually articulated [ʋ] paired with acoustic [b] is a viable McGurk viseme when paired with the appropriate sociolinguistic experience for listeners. Understanding whether this holds in other types of English dialects where /v/ and /w/ are merged or substituted for one another (e.g. German accented English) will be useful to tease apart the types of sociolinguistic and raciolinguistic pressures listeners are under.

As noted throughout this dissertation, the sociolinguistic pressures of not only *how* to use English in these tasks, but also, which *type* of English to use in these tasks varies across the two participant populations (Irvine, Gal & Kroskrity 2000, Lippi-Green 2011, Sailaja 2012, Craft et al. 2020, Wiltshire 2020). Many American English participants likely don't consider that they speak with an accent, or if they do, they likely perceive their American English accent as being a standard from which they assess deviations in Global Englishes. For Indian English participants, this reality is very different. Regardless of whether speakers think of their English as filtered through a variety of Indic and Dravidian L1's or as distinct phonologized dialect of English, speakers are aware that their dialect deviates from the production targets of British and American English dialects and is viewed as marked by speakers of these varieties. This opens various avenues, not only to pursue further research with the varieties under investigation here, but to continue the strategic use of illusory stimuli to probe the degree to which the accents that listeners perceive are the product of their own experience in addition to the bottom up signals around them.

Appendices

Appendix A: Intervention Script

OPEN – Andrew & Ameya standing side by side

AMEYA 1

Hi and thank you for participating in this experiment. We hope that you’ve enjoyed the experience thus far and hope that you will have lots of questions once you are finished with all the tasks we’re having you complete today.

ANDREW 1

You probably recognize us from the stimuli that you’ve been viewing thus far. We will also be in later blocks of the experiment but for now we’d like to take a short break to talk to you about sound waves and how they travel through the spaces around us. Make sure to pay attention since there will be questions about what we cover in this portion.

CLOSE SHOT OF AMEYA

AMEYA 2

Often, when there are physical changes within the world, information about a disturbance moves away from the source of the disturbance in all directions. The information travelling from a disturbance does so in the form of waves. Waves necessarily need two components, the first is an instance of disturbance or variation and the second is a medium through which the wave travels. Two common wave types are transverse waves and longitudinal waves. Transverse waves are often the easiest to visualize. With these waves, the motion of the wave moves perpendicular to the source of the disturbance. This is what you see, for example, when a whip is snapped. When a whip is snapped, the force exerted by the person snapping the whip causes the energy to move down the whip creating a wave that results in the sound of the whip's crack. This wave is transverse because the wave's energy moves outwards away from the person, perpendicular to the up and down motion of the hand cracking the whip.

CLOSE SHOT OF ANDREW

ANDREW 2

The other type of wave that we'll talk about in this video is the longitudinal wave. Longitudinal waves are different from transverse waves in that the motion of the wave travels in a parallel direction to the source of the disturbance. Sound is a type of longitudinal wave. But an easy visualization of longitudinal waves involves springs or Slinkys. If you visualize a slinky laying on its side, you can push one end of the slinky to pass a wave through to the other side. This results in the other end of the slinky moving in the direction of the push and then returning to its original state. As compared to the whip example the hand of the person pushing the slinky moves in parallel with the direction of the wave generated by the disturbance. This is the same thing that happens when someone speaks. Air molecules are displaced by the vibration of the vocal folds and the articulators in the mouth and the energy from that disturbance moves through air parallel to the disturbance until hitting your ear.

SIDE BY SIDE SHOT

AMEYA 3

Sound waves are sometimes called pressure waves and it is helpful to think about sound as changes in pressure. When sound waves move through the air, they compress and expand air molecules resulting in pockets of higher and lower pressure. When thinking about vocal fold vibration, the regular patterns of displacement generated by the vocal folds passes that vibration through the medium of the air.

ANDREW 3

These pockets of low and high pressure create the peaks and troughs that you probably envision when you think about what an image of a sound wave looks like on a computer or in a drawing, like this one. Here, the peaks would be points of compression and the troughs points of expansion. The height of these peaks and troughs correspond to the amplitude of the wave. We perceive amplitude as loudness, thus the greater the amplitude the greater the perceived loudness.

CLOSE SHOT OF AMEYA

AMEYA 4

We can also describe waves by the differences in the distance between their peaks. The distance from one peak to the other in a wave is the wavelength. When you consider how many wavelengths pass a single point in a given time, say one second, you can calculate the waves frequency. The frequency of a sound wave is what we perceive as the pitch of the sound: how high or low the wave sounds. Higher frequency waves will have a higher pitch and lower frequency waves will have a lower pitch. But a waves perceived pitch can also change if the source of the wave is in motion. Think of a duck on a pond. As the duck moves across the water it generates ripples or waves. The waves in front of the duck bunch up and create ripples that are closer together while the ripples trailing behind the duck are more spread out. The difference between how frequently a ripple from the moving duck passes a stationary point, depending on the direction of the duck, has an effect on the perceived frequency of the wave. This effect also happens with sound and changes how we perceive the sound associated with a moving disturbance.

CLOSE SHOT OF ANDREW

ANDREW 4

Think about when you hear an ambulance or a fire engine siren when they're driving down the street. We've all had the experience, when the ambulance or fire engine passes by, that the siren sounds lower in pitch while it's trailing off. However, when one of these vehicles is parked the siren always sounds like it's the same pitch. This is called the Doppler effect, or the effect on the frequency of a wave when the waves source is moving. In this example, as the vehicle with the siren gets closer to where you are, the pitch of the siren sounds higher and once it passes you the pitch lowers. This is because of how we perceive the effect of motion of the waves source on the frequency of the wave. Like the example with the duck, the sound of the siren gets bunched up as the vehicle approaches and gets spread out as the vehicle passes by. It's not that anything changes about the siren or the sound it emits. Those details are constant, but the movement of the vehicle creates the perceived difference that we experience.

SIDE BY SIDE SHOT

AMEYA 5

This effect isn't only limited to sound though. It is also present in waves of light and can tell astronomers how stars move throughout the galaxy. We hope that you’ve learned something in this brief aside and thank you for your attention.

ANDREW 5

Once this video completes you will be prompted to answer a question about what we’ve covered. Then, you will continue on to another block of the experiment. Thanks again.

Appendix B: Experiment 1 Questionnaire

Our experience with different languages and different dialects of English can influence the way we produce and perceive speech. Having background information on these experiences can be helpful to researchers who study language in interpreting their results. Please provide the following information:

Age:

Gender:

Language(s) spoken in home as a child:

Language(s) spoken at home:

Other languages that you speak:

Place of birth:

Please indicate where, and for how long, you have lived in locations other than your birthplace:

Do speakers of your **first language** say that you speak with an accent? If yes, what accent do they say you have?

Do speakers of **any other languages you speak** say that you speak with an accent? If so, what accent do they say you have?

Do you have any known speaking or hearing deficits? If yes, please explain:

Did you use wired or bluetooth headphones for this task?

Appendix C: Experiment 2 Questionnaire

Our experience with different languages and different dialects of English can influence the way we produce and perceive speech. Having background information on these experiences can be helpful to researchers who study language in interpreting their results. Please provide the following information:

Age:

Gender:

Language(s) spoken in home as a child:

Language(s) spoken at home:

Other languages that you speak:

Place of birth:

Please indicate where, and for how long, you have lived in locations other than your birthplace:

Do speakers of your **first language** say that you speak with an accent? If yes, what accent do they say you have?

Do speakers of **any other languages you speak** say that you speak with an accent? If so, what accent do they say you have?

Do you have any known speaking or hearing deficits? If yes, please explain:

What did you think this experiment was about?

Bibliography

Alsius, A., Paré, M. and Munhall, K.G., 2018. Forty years after hearing lips and seeing voices: The McGurk effect revisited. Multisensory Research, 31(1-2), pp.111-144.

Anwyl-Irvine, A.L., Massonié, J., Flitton, A., Kirkham, N.Z., Evershed, J.K. 2019. Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods.*

Babel M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics,* 40, 177–189.

Basu Mallick, D., F Magnotti, J. and S Beauchamp, M., 2015. Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic bulletin & review*, *22*, pp.1299-1307.

Beauchamp, M.S., Nath, A.R. and Pasalar, S., 2010. fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. Journal of Neuroscience, 30(7), pp.2414-2417.

Bernstein, L.E., Auer Jr, E.T. and Takayanagi, S., 2004. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1-4), pp.5-18.

Bertelson, P, Vroomen, J, de Gelder, B. 2003. Visual recalibration of auditory speech identification: A McGurk Aftereffect. *Psychological Science* 14(6). 592-597

Boersma, Paul & Weenink, David. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.4.01, retrieved 12 November 2022 from http://www.praat.org/

Bouvachith D., Calloway I., Craft J., Hildebrandt T., Tobin S., Beddor. 2019. Perceptual influences of social and linguistic priming are bidirectional. *Proceedings from the 2019 International Congress on Phonetic Sciences,* Melbourne, Australia.

Bradlow, A.R. and Bent, T. 2008. Perceptual adaptation to non-native speech. *Cognition*, *106*(2), pp.707-729.

Brancazio, L., Miller, J.L. and Paré, M.A. 2003. Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, *65*(4), pp.591-601.

Brumm, H., & Slabbekoorn, H. 2005. Acoustic Communication in Noise. *Advances in the Study of Behavior*, 151-209.

Burnham, D. 1998. Language specificity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech.* Hove, UK: Psychology Press. 27-60.

Burnham D., Dodd B. 2017. Language-General Auditory-Visual Speech Perception: Thai-English and Japanese-English McGurk Effects. *Multisensory Research.* (31). 79-110Chomsky N, Halle M. 1968. *The Sound Pattern of English.* New York: Harper Row

CIEFL. 1972. *The Sound System of Indian English*. Monograph 7. Hyderabad: CIEFL.

Coetzee, A.W., Beddor, P.S., Styler, W., Tobin, S., Bekker, I. and Wissing, D. 2022. Producing and perceiving socially structured coarticulation: Coarticulatory nasalization in Afrikaans. *Laboratory phonology*, *13*(1).

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. and Deltenre, P. 2002. Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. Clinical Neurophysiology, 113(4), pp.495-506.

Craft, J.T., Wright, K.E., Weissler, R.E. and Queen, R.M. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, *6*, pp.389-407.

Dahan, D., Drucker, S.J. and Scarborough, R.A. 2008. Talker adaptation in speech perception: Adjusting the signal or the representations?. *Cognition*, *108*(3), pp.710-718.

Diehl, R.L., Lotto, A.J. and Holt, L.L., 2004. Speech perception. Annu. Rev. Psychol., 55, pp.149-179.

Drager, K. 2010. Sociophonetic variation in speech perception. *Language and Linguistics Compass.* 4, 473-480.

Drager K., Kirtley M.J. 2016. Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. *Awareness and Control in Sociolinguistic Research*, ed. AM Babel, pp. 1–24. Cambridge, UK: Cambridge Univ. Press

Dufour, S., & Nguyen, N. 2013. How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*.

Eckert P. 2008. Variation and the indexical field. *Journal of Sociolinguistics*. 12(4):453–76

Eskelund K., MacDonald E.N., Andersen T.S. 2015. Face configuration affects speech perception: evidence from a McGurk mismatch negativity study. *Neuropsychologia.* 66:48–54.

Flege, J.E. and Bohn, O.S., 2021. The revised speech learning model (SLM-r). Second language speech learning: Theoretical and empirical progress, pp.3-83.

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

Fowler, C.A., 1986. An event approach to the study of speech perception from a direct–realist perspective. Journal of phonetics, 14(1), pp.3-28.

Fowler, C.A., 1996. Listeners do hear sounds, not tongues. The Journal of the Acoustical Society of America, 99(3), pp.1730-1741.

Fowler, C. A. 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396–413.

Fowler, C. A. 2004. Speech as a supramodal or amodal phenomenon. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 189–201). Cambridge, U.K.: Cambridge University Press.

Fowler, C.A., Dekle D.J. 1991. Listening with eye and hand: cross modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance.* 17(3). 816-828.

Fowler, C.A., & Iskarous, K. 2013. Speech Production and Perception. In I. B. Weiner, A. F. Healey, & R. W. Proctor (Eds.), *Experimental Psychology.* (2nd ed., Vol. 4, Handbook of Psychology, pp 236-264). Wiley.

Fuchs, R. 2019. Almost [w]anishing: The elusive /v/-/w/ contrast in educated Indian English. *Proceedings from the 2019 International Congress of Phonetic Sciences,* Melbourne3 Australia.

Gargesh, R. 2008. Indian English: phonology. *The handbook of varieties of English: vol.1, phonology*, ed. by Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie and Clive Upton, 993–1002. Berlin: Mouton de Gruyter.

Gentilucci, M., & Cattaneo, L. 2005. Automatic audiovisual integration in speech perception. *Experimental Brain Research*, *167*(1), 66–75.

Gick B., Derrick D. 2009. Aero-tactile integration in speech perception, *Nature*, 462(7272): 502–504

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

Goldstein, L. and Fowler, C.A., 2003. Articulatory phonology: A phonology for public language use. Phonetics and phonology in language comprehension and production: Differences and similarities, pp.159-207.

Green, K.P., Kuhl, P.K. and Meltzoff, A.N., 1988. Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *The Journal of the Acoustical Society of America*, *84* (S1), pp.S155-S155.

Green K.P., Kuhl P.K., Meltzoff A.N., Stevens E.B. 1991. Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics*. 50: 524–536. pmid:1780200

Green K.P., Kuhl P.K.. 1991. Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*. 17: 278–288. pmid:1826317

Green, K.P. and Gerdman, A., 1995. Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), p.1409.

Green, K.P. and Norrix, L.W., 1997. Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. *Journal of Speech, Language, and Hearing Research*, *40*(3), pp.646-665.

Hamann, Silke. 2003. *The Phonetics and Phonology of Retroflexes*. Ph.D. dissertation. Utrecht: LOT Press.

Hardison, D. M. 1999. Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect, *Language Learning.*, 49:213- 283.

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

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

Hayes, B.P., 2004. Phonetically Driven Phonology. Functionalism and Formalism in Linguistics, p.243.

Hickok, G., Rogalsky, C., Matchin, W., Basilakos, A., Cai, J., Pillay, S., Ferrill, M., Mickelsen, S., Anderson, S.W., Love, T. and Binder, J., 2018. Neural networks supporting audiovisual integration for speech: A large-scale lesion study. Cortex, 103, pp.360-371.

Hillenbrand, J.M. 2003. American English: Southern Michigan. *Journal of the International Phonetic Association*, *33*(1), pp.121-126.

Honorof, D. N., Weihing, J., & Fowler, C. A. 2011. Articulatory events are imitated under rapid shadowing. *Journal of Phonetics*, *39*(1), 18–38.

Irvine, J.T., Gal, S. and Kroskrity, P.V. 2009. Language ideology and linguistic differentiation. *Linguistic anthropology: A reader*, *1*, pp.402-434.

Ito T., Tiede M., Ostry D. J. 2009. Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of United States of America*. 106:1245–1248.

Jaeger TF, Weatherholtz K. 2016. What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology.* 7: 1115

Johnson K. 2006. Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *Journal of Phonetics.* 34:485–99

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).*

Kleinschmidt, D.F. 2019. Structure in talker variability: How much is there and how much can it help?. *Language, cognition and neuroscience*, *34*(1), pp.43-68.

Kraljic, T., Samuel, A.G. 2006. Generalization in perceptual learning. *Psychonomic Bulletin & Review.* 13 (2) 262 – 268

Kraljic, T., Samuel, A.G. and Brennan, S.E. 2008. First impressions and last resorts: How listeners adjust to speaker variability. *Psychological science*, *19*(4), pp.332-338.

Kutlu, E. 2023. Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different English varieties. *Journal of Multilingual and Multicultural Development*, *44*(6), pp.511-525.

Kwon, H. 2019. The role of native phonology in spontaneous imitation: Evidence from Seoul Korean.

Ladefoged, P., & Bhaskararao, P. 1983. Non-quantal aspects of consonant production: A study of retroflex consonants. *Journal of Phonetics*, *11*(3), 291–302.

Liberman, A.M. and Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition, 21(1), pp.1-36.

Lippi-Green R. 2011. English with an Accent: Language, Ideology and Discrimination in the United States. London: Routledge. 2nd ed.

Lombard, E. 1911. Le signe de l'élévation de la voix. *Annales des Maladies de L'Oreille et du Larynx.* 37, 101–119.

MacDonald, J. and McGurk, H. 1978. Visual influences on speech perception processes. *Perception & psychophysics*, *24*(3), pp.253-257.

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.

Marr, D. 1982. *Vision*. New York: W.H. Freeman

Mazzoni, D, Dannenberg R.D. 2002. A Fast Data Structure for Disk-Based Audio Editing. *Computer Music Journal*. 26 (2): 62–76.

McCullough, E. A., & Clopper, C. G. 2016. Perceptual subcategories within non-native English. *Journal of Phonetics*, *55*, 19–37.

McGowan KB. 2016. Sounding Chinese and listening Chinese: awareness and knowledge in the laboratory. In *Awareness and Control in Sociolinguistic Research*, ed.AM Babel, pp. 25–61. New York: Cambridge University Press

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

McMurray B, Tanenhaus MK, Aslin RN. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition,* 86: B33-B42

Mitterer, H. and Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), pp.168-173.

Mitterer, H., & Müsseler, J. 2013. Regional accent variation in the shadowing task: Evidence for a loose perception–action coupling in speech. *Attention, Perception, & Psychophysics*, *75*(3), 557–575.

Munhall KG, Gribble P, Sacco L, Ward M. 1996. Temporal constraints on the McGurk effect. *Perception & Psychophysics*. 58: 351–362. pmid:8935896

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

Nath, A.R. and Beauchamp, M.S., 2012. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage*, 59(1), pp.781-787.

Niedzielski N. 1999. The effect of social information on the perception of sociolinguistic variables. *J. Lang.Soc. Psychol.* 18(1):62–85

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

Norris, D., McQueen, J.M., Cutler, A. 2003. Perceptual learning in speech. *Cognitive Psychology*. 47(2) 204 – 238

Ohala, J.J., 1996. Speech perception is hearing sounds, not tongues. The Journal of the Acoustical Society of America, 99(3), pp.1718-1725.

Pardo, J. S. 2013. Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*.

Peirce CS, Hoopes JE. 1991. *Peirce on Signs: Writings on Semiotic*. Chapel Hill, NC: Univ. N.C. Press

Perkins, Katherine & Adams, Wendy & Dubson, Michael & Finkelstein, Noah & Reid, Sam & Wieman, C. & LeMaster, Ron. 2006. PhET: Interactive Simulations for Teaching and Learning Physics. The Physics Teacher. 44. 18-23. 10.1119/1.2150754.

Pierrehumbert J. 2002. Word-specific phonetics. *In Laboratory Phonology* 7, ed. C Gussenhoven, N Warner,pp. 101–39. Berlin: Mouton de Gruyter

Prince, A. and Smolensky, P., 2004. Optimality Theory: Constraint interaction in generative grammar. Optimality Theory in phonology: A reader, pp.1-71.

Redford, M. and Baese-Berk, M., 2023. Acoustic Theories of Speech Perception. In *Oxford Research Encyclopedia of Linguistics.*

Reinisch, E., Mitterer, H. 2016. Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*. 55: 96-108

Rosa, J.D. 2016. Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology*, *26*(2), pp.162-183.

Rosa, J. and Flores, N. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, *46*(5), pp.621-647.

Rosenblum, L. D., & Saldaña, H. M. 1996. An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 22(2), 318–331.

Rosenblum, L. D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 51–78). Malden, MA: Blackwell.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. 2007. Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science*, 18, 392.

Rosenblum, L. D., Dorsi, J., & Dias, J.W. 2016. The impact and status of Carol Fowler'ssupramodal theory of multisensory speech perception. Ecological Psychology, 28, 262– 294.

Rosenblum, L. D., Dias, J.W., & Dorsi, J. (2017). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology,* 28, 1–23

Rosenblum, L., 2019. Audiovisual speech perception and the McGurk effect. *Oxford Research Encyclopedia, Linguistics*.

Russell, DA. 2014. The Doppler Effect and Sonic Booms. Retrieved from: https://www.acs.psu.edu/drussell/Demos/doppler/doppler.html

Sailaja, P., 2012. Indian English: Features and sociolinguistic aspects. *Language and Linguistics Compass*, *6*(6), pp.359-370.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W. and Foxe, J.J. 2007. Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, *45*(3), pp.587-597.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.T. and Simola, J. 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience letters*, *127*(1), pp.141-145.

Samuel, A.G. and Kraljic, T., 2009. Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), pp.1207-1218.

Schertz, J. and Paquette-Smith, M., 2023. Convergence to shortened and lengthened voice onset time in an imitation task. JASA Express Letters, 3(2).

Sekiyama, K., and Tohkura, Y. 1991. McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*. 90, 1797 1805.

Sekiyama K., and Burnham D. 2008. “ Impact of language on development of auditory-visual speech perception,” *Developmental Science*. 11(2), 306–320.10.1111/j.1467-7687.2008.00677

Shams, L. 2011. Early integration and Bayesian causal inference in multisensory perception. In M. M. Murray & M. T. Wallace (Eds.), The neural bases of multisensory processes (pp. 217–232). Boca Raton, FL: CRC Press.

Shigeno, S. 2002. Anchoring effects in audiovisual speech perception. *Journal of the Acoustical Society of America*,111, 2853–2861.

Shockley, K., Sabadini, L., & Fowler, C. A. 2004. Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

Silverstein M. 2003. Indexical order and the dialectics of sociolinguistic life. *Lang. Commun*. 23(3–4):193–229

Sirsa, H., & Redford, M. A. 2013. The effects of native language on Indian English sounds and timing patterns. *Journal of Phonetics*, *41*(6), 393–406.

Sonntag, S.K. 2009. Linguistic globalization and the call center industry: Imperialism, hegemony or cosmopolitanism?. *Language Policy*, *8*, pp.5-25.

Staum-Casasanto L. 2008. Does social information influence sentence processing? *Proceedings of the Annual Meeting of the Cognitive Science Society*. 30(30):799–804

Strand E.A, Johnson K. 1996. Gradient and visual speaker normalization in the perception of fricatives. In *Natural Language Processing and Speech Technology. Results of the 3rd KOVENS Conference, Bielefeld, October, 1996*, ed D Gibbon, pp. 14-26. Berlin: Mouton de Gruyter

Strand EA. 1999. Uncovering the role of gender stereotypes in speech perception. *J. Lang. Soc. Psychol*. 18(1):86–99

Stevens, K.N., 1989. On the quantal nature of speech. Journal of phonetics, 17(1), pp.3-45.

Sumby, W.H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), pp.212-215.

Sumner M, Kim SK, King E,McGowan KB. 2014. The socially-weighted encoding of spoken words: a dual route approach to speech perception. *Frontiers in Psychology*. 4:1015

Tiippana, K., Andersen, T. S., & Sams, M. 2004. Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology,* 16(3), 457–472.

Tiippana K. 2014. What is the McGurk effect? *Frontiers in Psychology*. 5: 725. pmid:25071686

Trude, A.M., Brown-Schmidt, S. 2012. Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*. 27(7/8). 979-1001.

van Wassenhove, V., Grant, K. W., and Poeppel, D. 2007. Temporal window of integration in auditory visual speech perception. *Neuropsychologia* 45, 598–607.

Wade, L., 2022. Experimental evidence for expectation-driven linguistic convergence. *Language*, *98*(1), pp.63-97.

Walker S., Bruce V., O'Malley, C. 1995. Facial identity and facial speech processing: Familiar faces and voices in the McGurk Effect. *Perception and Psychophysics.* 57(8). 1124-1133.

Wang, Y., Behne, D.M. and Jiang, H., 2008. Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), pp.1716-1726.

Whalen, D. H. 1984. Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics* 35, 49-64.

Wiltshire, C. R. 2005. The "Indian English" of Tibeto-Burman language speakers. *English World-Wide. A Journal of Varieties of English*, *26*(3), 275–300.

Wiltshire, C. R., & Harnsberger, J. D. (2006). The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. *World Englishes*, *25*(1), 91–104.

Wiltshire, C.R., 2020. *Uniformity and variability in the Indian English accent*. Cambridge University Press.

Zhao, S.Y., 2010. Stop-like modification of the dental fricative/ð: An acoustic analysis. *The Journal of the Acoustical Society of America*, *128*(4), pp.2009-2020.