GLOSSA psycholinguistics

Aoki, N. B., & Zellou, G. (2025). When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception. *Glossa Psycholinguistics*, 4(1): 10, pp. 1–47. DOI: https://doi.org/10.5070/G601121217

eScholarship University of California

When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception

Nicholas B. Aoki, Department of Linguistics, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, California, USA 95616, nbaoki@ucdavis.edu

Georgia Zellou, Department of Linguistics, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, California, USA 95616, gzellou@ucdavis.edu

Sociolinguistic research finds that: (i) the speech signal contains talker-specific and socio-indexical structure, with talkers varying idiosyncratically within the same social category and systematically across categories; (ii) both talker-specific and socio-indexical variation influence speech perception. What is unclear is how sociolinguistic perception arises – following exposure to an unfamiliar, socially-mediated variant, how do listeners learn that this feature is characteristic of a broader social group and can generalize to other group members? The current study exposed listeners to an unattested variant in L1-English (a /p/ to [b] phonetic shift), investigating how the number of exposure talkers mediates cross-talker generalization. All participants completed an exposure phase (phrase-final keyword identification) followed by a test phase (categorization along a *buy-pie* continuum for a novel female and male talker in separate blocks). Experiment 1 exposed listeners to a single shifted female talker ("The novel is now in *brint*") and a single unshifted male talker. Experiment 2 presented two shifted female and two shifted male talkers. We find: (i) no generalization in Experiment 1 (no difference in *buy-pie* response between the novel talkers); (ii) robust generalization in Experiment 2 (greater *pie* response for the novel female than the novel male talker), but only when the novel female block is presented first (i.e., generalization is short-lived). Taken together, the results support a *numerosity account*: when a previously unheard social variant is presented, multiple talkers per social group seem to be necessary for socially-mediated, cross-talker generalization. This study highlights a critical role of the listener's social experiences on generalization – multi-talker exposure might be unnecessary when exposed to more familiar types of speech (e.g., L2-accented English) and necessary when exposed to completely unfamiliar variants. Overall, the present experiments enhance our theoretical understanding of cross-talker generalization and offer insights into the emergence of sociolinguistic perception.

1. Introduction

Imagine that you are traveling to present a poster in a country you have never been to before. You get off the plane, and as you walk through the airport terminal, you trip and drop your belongings. A woman quickly approaches you and says, “Let me get your *boster* for you”. Interestingly, this stranger has produced the initial consonant of ‘poster’ with a token that is acoustically [b]-like in your own dialect of English. You have a short conversation with this woman and realize that her dialect is essentially the same as yours, except that all /p/-initial words begin with an acoustically [b]-like segment (‘paint’ = *baint*, ‘pulse’ = *bulse*, etc.). Given that this is the first person you have met in this country, you might wonder about the source of this /p/ to [b] phonetic shift. Is the shift limited to this particular individual, or is it a systematic characteristic of a broader social group, such as all women within this country?

Real-life registers and dialects do not consist of singular phonetic shifts (Wolfram & Schilling, 2015). Yet even this simplified scenario alludes to two important observations that highlight the complexity of adapting to novel talkers. First, the speech signal simultaneously contains both *talker-specific covariation* and *socio-indexical covariation* – linguistic variants can correlate with both individual talkers (Chodroff & Wilson, 2017; Yu & Zellou, 2019) and broader social categories (Carignan & Zellou, 2023; Labov, 1966; Zellou & Tamminga, 2014). Second, the speech signal often contains ambiguity (Kleinschmidt & Jaeger, 2015). In particular, it may initially be unclear whether an unfamiliar variant is an idiosyncrasy associated with an individual talker or a socially meaningful pattern associated with a group of talkers.

Extensive research suggests that, given sufficient experience, listeners learn which variants are correlated with social cues, and they can leverage their knowledge of socio-indexical covariation to guide perception of the speech signal (Campbell-Kibler, 2010). In a classic study, Strand and Johnson (1996) presented L1-English listeners with an acoustically gender-ambiguous voice and showed an image of either a stereotypically male or female face. Listeners were asked to categorize tokens along a *sod-shod* continuum, and they provided more *sod* responses when the voice was paired with a male face than with a female face. Participants therefore recognize that /s/- and /ʃ/-variation is gender-mediated, not purely idiosyncratic, and they *generalize* their prior knowledge to a novel speaker based on (visual) cues to apparent gender. Similar types of perceptual effects have been found for many other social categories, such as age (Walker & Hay, 2011), nationality (Niedzielski, 1999), dialect (D’Onofrio, 2015), ethnicity (Babel & Russell, 2015), and apparent humanity (Aoki et al., 2022).

Studies of sociolinguistic perception often presume that participants already possess ample experience listening to the linguistic variants that are presented. An understudied question, however, is how knowledge of socio-indexical covariation emerges in the first place. Returning to the anecdote outlined earlier, imagine that a /p/ to [b] phonetic shift is truly gender-mediated in a hypothetical English dialect, not just produced by a single talker. How do listeners learn that

this acoustic feature represents socio-indexical covariation, and when do they *generalize* this shift to their expectations about novel speakers? Although extensive research has been conducted on perceptual adaptation of speaker idiosyncrasies (Norris et al., 2003; Tzeng et al., 2021) and L2-accented speech (Bradlow & Bent, 2008; Xie et al., 2021), relatively little work explicitly frames perceptual adaptation as a question of sociolinguistic perception (cf. Aoki & Zellou, 2023a; Kleinschmidt et al., 2018; Zellou et al., 2023). The current study begins to address this gap by examining how cross-talker generalization is affected by the number of exposure talkers and the socio-indexical structure of exposure.

The rest of the introduction is structured as follows. 1.1 reviews lexically guided perceptual learning and the ideal adapter framework, which serve as the experimental approach and theoretical framing of the current study, respectively. 1.2 revisits the debate about how the number of the exposure talkers affects generalization, offering a sociolinguistic explanation to account for conflicting findings. 1.3 explains why this work specifically examines covariation between speaker gender and stop consonant production in American English. Finally, 1.4 delineates the study design and hypotheses.

1.1 Review of lexically guided perceptual learning and the ideal adapter framework

Work on lexically guided perceptual learning has demonstrated that listeners can readily adapt to idiosyncratic phonetic shifts produced by individual talkers (Samuel & Kraljic, 2009). Participants in a typical adaptation experiment complete an exposure phase (often a lexical decision task) followed by a test phase (usually a categorization task). When listeners are placed in a talker-specific condition (i.e., the same speaker is presented at both exposure and test), test phase categorization is altered based on the lexical bias in exposure. Consistently replacing /d/ with an ambiguous segment between /d/ and /t/ in exposure (e.g., ‘croco?ile’) results in more /ada/ response along an /ada/-/ata/ continuum at test, while replacing /t/ with the same ambiguous segment (e.g., ‘fron?ier’) leads to a greater /ata/ response. Talker-specific adaptation is a highly replicated phenomenon that underscores the flexible nature of the perceptual system (Cummings & Theodore, 2023; Norris et al., 2003; Zellou et al., 2023).

Perceptual learning can be neatly accounted for through the ideal adapter framework (Kleinschmidt & Jaeger, 2015). An ideal adapter leverages their prior knowledge about acoustic cue distributions to make the most accurate predictions possible about the incoming speech signal. Importantly, these predictions can be updated through experience, allowing the listener to adjust to novel input in the face of high acoustic variability. For example, after exposure to a talker who repeatedly produces /d/ as an ambiguous segment between /d/ and /t/ (e.g., ‘croco?ile’, ‘legen?ary’, etc.), listeners can then develop a talker-specific mental model. The model would predict that segments with ambiguous acoustic cues (e.g., voice onset time between

prototypical /d/ and /t/) are more likely to be classified as /d/ for this talker compared to the general population. Listeners can then utilize this talker-specific model to help them make future predictions (e.g., that an ambiguous token between /ada/ and /ata/ is more likely to be /ada/ when produced by this particular talker).

Unlike talker-specific adaptation, cross-talker generalization is much more inconsistent (i.e., after exposure to a phonetic shift or accent, listeners may or may not be willing to apply what they have learned to a *novel* talker; Weatherholtz & Jaeger, 2016). The ideal adapter framework can explain this relative restraint as striking a balance between being efficient while also maintaining accurate predictions (Kleinschmidt & Jaeger, 2015). On the one hand, if listeners are initially exposed to a talker with a phonetic shift and then hear a novel test talker who likely produces the same shift, it would be more efficient to generalize – with little decrease in prediction accuracy, the same mental model can be applied to both talkers without expending any extra effort in creating a new model. Overgeneralizing, however, can lead to incorrect predictions if the test talker is unlikely to produce the same phonetic shift. Taken together, the ideal adapter framework predicts that listeners somehow formulate expectations about the relationship between exposure talkers and novel talkers, and that these beliefs then modulate cross-talker generalization.

Although it is uncontroversial that cross-talker generalization is constrained by the relationship between the properties of the exposure and test phases, the specific constraints are still under debate (Baese-Berk et al., 2020). The current study manipulates one type of well-studied constraint (the number of exposure talkers), but re-examines prior work from a sociolinguistic perspective.

1.2 Taking a sociolinguistic perspective: How does the number of exposure talkers affect cross-talker generalization?

Bradlow and Bent (2008) proposed that multiple exposure talkers are necessary for cross-talker generalization in perceptual adaptation to L2-accents. Relative to a control condition (exposure to L1-English speakers), transcription accuracy in noise for a novel Mandarin-accented English speaker was only facilitated for listeners with recent exposure to *multiple* Mandarin-accented talkers, not for participants who had previously heard a *single* Mandarin-accented talker. Kleinschmidt and Jaeger (2015) attributed this effect to the distinct mental models that listeners develop following single-talker and multi-talker exposure. On the one hand, participants who are exposed to just one talker might assume that the particularities of the accent are idiosyncrasies, thereby resulting in a talker-specific model that does not generalize to novel talkers. However, listeners who are exposed to multiple talkers with the same L2-accent can more readily recognize the covariation in the speech signal between category membership (Mandarin-accented English speaker) and acoustic variation (e.g., devoicing of word-final stop consonants). This recognition ostensibly leads to the development of a talker-general mental model for Mandarin-accented

English, which then gives rise to successful generalization (i.e., enhanced comprehension of a novel talker with the same, Mandarin-accented English accent).

In contrast to Bradlow and Bent (2008), however, more recent work has found no effect of the number of exposure talkers on generalization (Xie & Myers, 2017; Xie et al., 2021). Participants in Xie and Myers (2017) completed an exposure phase (auditory lexical decision task) followed by a test phase (cross-modal priming task). Everyone in the experiment was tested on the same (novel) Mandarin-accented English speaker. In the critical conditions, listeners were either exposed to multiple Mandarin-accented English speakers, a single Mandarin-accented English speaker that was acoustically similar to the novel test talker, or to a single Mandarin-accented English speaker that was acoustically different from the test talker. Two major findings emerged: (i) both the multi-talker exposure and the 'single-talker, acoustically similar' conditions resulted in cross-talker generalization, with no meaningful difference observed between the two conditions; (ii) generalization was blocked in the 'single-talker, acoustically different' condition. Xie and Myers (2017) concluded that acoustic similarity between exposure and test talkers mediates generalization, not simply the number of exposure talkers. Multi-talker exposure only leads to generalization when one of the exposure talkers happens to be acoustically similar to the novel speaker.

Providing further evidence against the primacy of the number of exposure talkers, Xie et al. (2021) conducted a replication study of Bradlow and Bent (2008) and did not find a meaningful difference in transcription accuracy between the critical single- and multi-talker exposure conditions. Xie et al. (2021) attribute this lack of replication to "the removal of the design confound [in Bradlow and Bent (2008)] coupled with increased statistical power" (p. e37), where the "design confound" refers to "the single- and multitalker conditions employ[ing] different L2 exposure talkers" (p. e36). In other words, the lack of generalization in the single-talker exposure condition of Bradlow and Bent (2008) may have merely occurred because the authors coincidentally selected exposure talkers that were all acoustically different from the test talker (and conversely, the generalization in the multi-talker exposure condition might have been blocked if none of the exposure talkers were acoustically similar to the test talker).

While not downplaying methodological considerations, there could additionally be a social explanation for why the number of exposure talkers has not affected generalization in recent work. Xie and Myers (2017) and Xie et al. (2021) both examine Mandarin-accented English, and in general, the proportion of L2-accented English speakers is rising in the United States (Graddol, 2003; ShareAmerica, 2023). Although Xie and Myers (2017) recruited "monolingual English speakers with…no or minimal prior experience with Mandarin-accented English or the Mandarin language" (p. 33) and Xie et al. (2021) "excluded participants from analysis who reported a high degree of familiarity with Chinese or Chinese-accented English" (p. e27), many listeners likely had some experience with other L2-English accents.

L2-English accents often share acoustic properties, such as a slower speaking rate (Baese-Berk & Morrill, 2015) and reduced usage of spectral cues when producing tense/lax vowel contrasts (Feng & Wang, 2024; Sidaras et al., 2009). Therefore, despite limited self-reported experience with Mandarin-accented English, listeners in Xie and Myers (2017) and Xie et al. (2021) could have initially engaged in accent-independent adaptation (Baese-Berk et al., 2013), having recognized certain features in the exposure phase (e.g., devoiced word-final stop consonants are features of both Mandarin-accented and Dutch-accented English; Eisner et al., 2013; Xie & Fowler, 2013). After this head start, participants may have attuned more to the specific acoustic properties of the exposure talker, with generalization only occurring when the exposure and test talkers were sufficiently similar acoustically.

Given their likely prior experience with L2-English accents, listeners in Xie and Myers (2017) and Xie et al. (2021) presumably did not consider the acoustic properties in exposure to be entirely idiosyncratic, which could make the number of exposure talkers irrelevant. The distinction between single- and multi-talker exposure conditions might only be important when listeners are truly unfamiliar with a variant (e.g., the /p/ to [b] shift discussed in Section 1, where a novel L1-English speaker produces words like 'poster' as *boster*). A novel phonetic shift might be initially thought of as idiosyncratic if it is heard in only one talker. To prove that a variant is socially-mediated and generalizable to novel talkers, it could be necessary to hear the shift from multiple talkers within the same social group.

Thus, it is still unclear whether multiple exposure talkers are *necessary* for cross-talker generalization. L2-accented speech may not be the most appropriate test case for addressing this question, given the ubiquity of L2-accented speakers. An ideal test case could be a variant that is likely to be treated as an idiosyncrasy when produced by one speaker and as a socially-mediated feature when produced by more than one speaker from an identifiable social group.

1.3 Motivation of covarying gender and stop consonant production

The current study examines the covariation of (binary) gender and stop consonant production in L1-accented English, with females producing /p/ as [b] and males producing prototypical /p/.[1] There are several reasons for employing this specific test case. For one, gender is a highly salient and recognizable socio-indexical cue (Barreda & Predeck, 2024; Eckert, 1989). Male and female voices are often acoustically distinct (Barreda, 2021), and listeners can use the (on average) lower f0 and formant frequencies of males to identify speaker gender with high confidence and near-ceiling accuracy (Hillenbrand & Clark, 2009). Listeners also know, consciously or not, that gender can be relevant for speech perception in certain cases (e.g., an ambiguous sound between /s/ and /ʃ/ is more likely to be categorized as /s/ if the talker is male; Yu, 2010).

[1] The current test case revolves around binary gender. However, gender is much more complex than a male-female dichotomy – many individuals identify as non-binary (Eckert, 2014), and gender intersects with other social categories (Calder & King, 2022). Expanding the present study beyond binary gender is an important step for future work.

Critically, however, compared to other phonological contrasts, such as fricatives and vowels, comparatively little covariation currently exists between gender and American English stop consonant production (Kleinschmidt, 2019; Morris et al., 2008). Among studies that do find gender effects (e.g., Robb et al., 2005; Swartz, 1992), the result is almost always the exact opposite of the covariation we have constructed for the current study – male speakers usually produce voiceless consonants with a shorter voice onset time than female speakers (i.e., it is *male* speakers who produce /p/ as a more acoustically [b]-like sound). A further distinctive feature of the current test case is the use of a "bad map", or a sound that "fall[s] unambiguously into an unintended category" (Sumner, 2011, p. 132). Shifts from /p/ to [b] can be heard in English, but usually in L2-accented English (Flege & Eefting, 1987; Solé, 2018), not L1-accented English.

In summary, the current study correlates a highly salient variant (a "bad map" /p/ to [b] shift) with a highly salient social cue (speaker gender) in a way that listeners have likely never experienced in L1-English. Unlike L2-accented English, which has been the focus of many prior experiments on adaptation (Xie & Myers, 2017; Xie et al., 2021), there is a greater chance that the phonetic shift in this study will be truly regarded as an idiosyncrasy upon initial exposure. The current study can therefore more effectively investigate how the number of exposure talkers affects cross-talker generalization, while additionally lending insight into the *emergence* of sociolinguistic perception.

1.4 The current study and hypotheses

Two experiments were conducted to examine how the number of exposure talkers and the socio-indexical structure of exposure influence cross-talker generalization. All participants completed an exposure phase (identification of phrase-final keywords; e.g., "The wall needed a new coat of *paint*") followed by a test phase (categorization of stimuli along a *buy-pie* continuum for both a novel female and novel male speaker). Within each experiment, there were two types of exposure conditions that varied in socio-indexical structure: (i) a critical condition with a gender-mediated phonetic shift (/p/ produced as [b]; e.g., 'paint' as *baint*); (ii) a control condition with no shifted keywords. Across experiments, the critical conditions differed in the number of exposure talkers – whereas Experiment 1 only presented one female and one male talker, Experiment 2 exposed listeners to two female and two male talkers. The key question is whether the experimental manipulations induce cross-talker generalization. When comparing the control and critical conditions, is there a difference in *buy-pie* categorization for the novel test talkers?[2]

The present study adjudicates between three possible accounts of perceptual adaptation: (i) talker normalization; (ii) sufficient similarity; (iii) numerosity. The predictions of each account are summarized in **Table 1** and further explained in 1.4.1–1.4.3.

[2] Xie et al. (2023) have recently pointed out that lexically guided perceptual learning can potentially be accounted for by at least three different mechanisms: (i) low-level normalization; (ii) shifts in category representations; (iii) changes in decision-making processes. The current study does not tease apart these possibilities and, thus, remains agnostic as to the underlying mechanisms behind any observed test phase categorization shift.

Table 1: Summary of predictions for each account. Each prediction ('No difference' or 'More *pie* response') compares the novel female talker to the novel male talker in the test phase of the critical condition. (In the control condition, neither exposure talker produces the /p/ to [b] phonetic shift, so *buy-pie* categorization is expected to be the same for both test talkers.) The cells immediately below the experiment titles refer to the structure of exposure in the critical condition, where F and M refer to 'female' and 'male', respectively.

Theoretical Account	**Experiment 1**	**Experiment 2**
	Shifted: 1 F **Unshifted: 1 M**	**Shifted: 2 F** **Unshifted: 2 M**
Talker Normalization	No difference	No difference
Sufficient Similarity	More *pie* response	More *pie* response
Numerosity	No difference	More *pie* response

1.4.1 Talker normalization

A *talker normalization account* proposes that listeners will 'normalize', or disregard, speaker gender differences in the exposure phase, resulting in similar responses for the novel female and male test talkers across all conditions (Joos, 1948; Liberman & Mattingly, 1985).[3] Given that the /p/ to [b] phonetic shift involves stop consonants, and stop consonant production does not strongly covary with gender in American English (Kleinschmidt, 2019), participants might ignore gender altogether (cf. Experiment 1 of Kraljic & Samuel, 2007). Besides gender, all of the critical exposure conditions are the same regarding the phonetic shift, since all conditions contain an equal number of shifted and prototypical tokens (cf. Experiment 2 of Tzeng et al., 2021). If talker gender is disregarded and, thus, the experiences hearing shifted and prototypical tokens are all weighted equally, then test phase categorization should be the same across all experiments. In other words, there should be no meaningful differences in the proportion of *pie* responses between the novel female and male talkers in the critical conditions.

1.4.2 Sufficient similarity

Unlike a talker normalization account, sufficient similarity and numerosity accounts both assume that speaker gender will impact adaptation. However, the latter two accounts diverge in their predictions about how the number of exposure talkers might affect the results. A *sufficient*

[3] Note that the term 'normalization' has been employed in many different ways in the speech perception literature. Here, normalization is intended to hearken back to more traditional theories which claimed that "variants of a particular type of linguistic unit are *normalized* [emphasis added] to arrive at an abstract, prototypical representation" (Nygaard, 2005, p. 391). More modern theories of normalization do not necessarily assume that social information is stripped away from the speech signal (Barreda, 2020).

similarity account claims that generalization should be triggered for any test talker who is sufficiently similar to the phonetically shifted exposure talkers, regardless of how many exposure talkers are heard (Xie & Myers, 2017; Xie et al., 2021). In the critical conditions, the /p/ to [b] phonetic shift is always associated with female speech in the exposure phase – thus, listeners in both experiments are expected to generalize the shift to the novel female test talker, given similarity in social category (gender) and in acoustics (e.g., higher f0 and formant frequencies; Hillenbrand & Clark, 2009). Generalization would specifically be realized as a greater proportion of *pie* responses for the novel female test talker than for the novel male test talker (i.e., if a speaker is expected to produce a /p/ to [b] shift, then more acoustically [b]-like tokens should be categorized as /p/ along a *buy-pie* continuum; see Sumner, 2011).

1.4.3 Numerosity

A *numerosity account* predicts that the number of exposure talkers should have an impact on the results, with generalization only occurring when listeners hear multiple talkers per gender (Bradlow & Bent, 2008). Given their lack of experience with the particular /p/ to [b] shift under investigation (see 1.3 for details), participants might treat the shift as a talker-specific idiosyncrasy when produced by only one talker, thus blocking generalization in Experiment 1 and leading to no differences in *pie* response between the novel test talkers. Listeners should only generalize in the critical condition of Experiment 2, since there is evidence that the shift is produced by a group of similar talkers (two female speakers) and is not a talker-specific idiosyncrasy. Generalization should specifically occur towards the novel female test talker, who aligns more closely with the phonetically shifted exposure speakers in terms of social and acoustic similarity.

2. Experiment 1: Exposure to a single talker per gender

Experiment 1 first exposed listeners to one male and one female speaker, and then presented a categorization task at test for a novel male and a novel female speaker. Listeners were either assigned to a critical condition (where the exposure female produced /p/ as [b] and the male remained unshifted) or to a control condition (where both exposure talkers were unshifted). Whereas a sufficient similarity account predicts a greater proportion of *pie* responses for the novel female talker than the novel male talker in the test phase, both talker normalization and numerosity accounts predict no difference in *pie* response between the novel talkers (see 1.4 for a more thorough explanation).

2.1 Methods

2.1.1 Stimuli

48 semantically predictable sentences were constructed for the exposure phase (see Appendix 1 for the full list). There were 32 critical stimuli and 16 filler stimuli (i.e., two-thirds critical and

one-third filler, following the ratio used in Lai and Tamminga (2024)), with all sentences containing a monosyllabic target word in phrase-final position. The filler sentences did not have any words with /b/ or /p/ (e.g., "She came out to the lake for a *swim*"). The critical sentences were also designed to avoid words with /b/ or /p/, except for the target words.[4] The critical target words contained one instance of /p/ (all word-initial), did not contain /b/, and were not part of a /b/-/p/ minimal pair (e.g., *paint*/**baint*; "The wall needed a new coat of *paint*").

All of the exposure sentences were automatically generated using neural text-to-speech (TTS) synthesis (Aoki et al., 2022; Aoki & Zellou, 2023b; Zellou et al., 2023), a highly naturalistic generation method (as opposed to concatenative TTS; Cohn & Zellou, 2020).[5] TTS voices were used in the current study because relative to recording speech in a laboratory setting, generating synthetic speech offers a high degree of control over the acoustic properties of each stimulus (e.g., every time 'paint' is generated with the Joanna voice, the voice onset time of the initial /p/ is always around 53 seconds). Extensive intra-speaker acoustic variation is well-documented using traditional recording methods (Aoki & Zellou, 2023c; Vonessen et al., 2024), and given that stimulus acoustics can affect perceptual adaptation (Lai & Tamminga, 2024; Xie & Myers, 2017), using controlled synthetic speech should make the results more reliable and replicable using the same methodology.

The test stimuli consisted of a 5-step *buy-pie* continuum for each of the six voices (see Appendix 2 for details about how the test stimuli were developed and normed). The exposure sentences were generated in three female (Joanna, Ruth, Salli) and three male (Joey, Matthew, Stephen) US-English voices from Amazon Polly. For each of the six voices, 16 filler sentences

[4] It should be mentioned that in one of the stimuli, there is a word containing /b/ ("The soccer team won the match *by* one point"; refer to the 21st row of Table 10 in Appendix 1). However, there are three reasons why this erroneous word is considered a relatively minor issue: (i) all listeners heard this instance of /b/, which does not explain the difference in results across experiments; (ii) the word *by* represents only 0.2% of the words listeners heard in the exposure phase (1/401); (iii) hearing canonical /b/ reinforces the mapping from [b] to /p/ in the shifted speakers (a more grave error would have been including a non-target word with canonical /p/, as in "*Paris* is a crowded place", since this could have blocked adaptation; Tzeng et al., 2021). Nevertheless, the intention was to avoid any instances of /b/ in the exposure stimuli, following prior work (Zellou et al., 2023), and the erroneous stimulus is noted here for transparency.

[5] Note that the use of synthetic voices, as opposed to naturally produced voices, is not expected to block effects of gender or other social attributes on perceptual adaptation. According to the Computers Are Social Actors (CASA) framework, if human-like qualities are observed (e.g., the use of language, as in the exposure phase of all experiments in the current study), then users often treat technological agents as humans, applying the same social heuristics and stereotypes (Nass & Moon, 2000; cf. Gambino et al., 2020). Indeed, synthetic speech has been successfully used in perceptual adaptation experiments (Maye et al., 2008; Zellou et al., 2023). Socially mediated linguistic behavior has also been observed with synthetic voices for a variety of other phenomena and tasks, including phonetic imitation (Zellou et al., 2021), apparent race judgments (Holliday, 2023), and judgments of perceived credibility (Pycha & Zellou, 2024). Overall, it is predicted that the results of the current study are replicable with naturally-produced voices, although future work should explicitly test this claim.

and two versions of the critical sentences (shifted and unshifted) were typed into the Amazon Web Services console (i.e., 6 voices * 16 filler + 6 voices * 32 critical * 2 versions = 96 + 384 = 480 stimuli total). All of the sentences were entered directly into the console and then downloaded. The only difference between the two versions of the critical sentences was that in the shifted version, the /p/-initial target words were changed orthographically to begin with *b* (e.g., 'paint' was typed as *baint*). All downloaded sentences were converted from .mp3 to .wav in the command line using FFmpeg (Tomar, 2006) and set to a presentation level of 60 dB SPL in Praat (Boersma & Weenink, 2021).

A brief acoustic analysis was conducted on both the exposure and test stimuli through Bonferroni-corrected paired t-tests. The average voice onset time (VOT) of word-initial stop consonants is displayed in **Table 2** for each speaker. Consistent with prior studies (Robb et al., 2005; Swartz, 1992), VOT of the unshifted /p/-initial stimuli was higher for each female speaker compared to each male speaker ($p < 0.001$ for nearly all comparisons; the only exception was the comparison between Joanna and Matthew, where $p = 0.01$). VOT of the unshifted /p/-initial stimuli was higher than that of the shifted /p/ to [b] for all female speakers in the exposure phase (all $p < 0.001$), in alignment with past work (Chodroff & Wilson, 2017). No statistically significant VOT differences were observed for the test phase continua between any of the speakers after using the Bonferroni correction (all $p > 0.006$).

A custom-made Praat script measured the mean f0 over the entire duration of all exposure and test phase productions (Cohn et al., 2021). The average f0 for each speaker is shown in **Table 3** and confirms that all three female speakers have a higher f0 than each of the male speakers in both the exposure and test phases (all $p < 0.001$).

Table 2: Average voice onset time (ms) of each speaker for word-initial stop consonants in unshifted exposure stimuli (e.g., 'paint'), shifted /p/ to [b] exposure stimuli (e.g., 'paint' as *baint*), and test phase stimuli (a 5-step *buy-pie* continuum). Dashes indicate speakers who did not produce shifted stimuli in the current study. The second, third, and fourth columns correspond to the female speakers (Joanna, Ruth, Salli), while the last three columns correspond to the male speakers (Joey, Matthew, Stephen).

	Joanna	**Ruth**	**Salli**	**Joey**	**Matthew**	**Stephen**
Exposure (Unshifted /p/)	59.66	67.60	68.05	53.45	54.29	44.72
Exposure (/p/ to [b])	21.31	20.68	19.93	-----	-----	-----
Test	37.71	39.78	36.75	36.11	39.80	37.45

Table 3: Average f0 (Hz) of each speaker for both the exposure and test phase stimuli. The second, third, and fourth columns correspond to the female speakers (Joanna, Ruth, Salli), while the last three columns correspond to the male speakers (Joey, Matthew, Stephen).

	Joanna	**Ruth**	**Salli**	**Joey**	**Matthew**	**Stephen**
Exposure	184.36	195.75	191.24	102.43	103.49	113.89
Test	172.93	217.80	220.27	102.12	107.84	95.17

2.1.2 Participants

All participants were recruited from Prolific, an online crowdsourcing platform. The Prolific demographic filters were used to narrow down the subject pool to individuals who were living in the United States, between 18 and 35 years old (inclusive), and whose reported first language was English. 419 participants from this restricted pool were recruited and compensated with $9 per hour ($1.80 for a 12-minute study). All subjects provided informed consent at the outset of the study, which received approval from the Institutional Review Board at the University of California, Davis.

Anyone who self-reported a hearing difficulty (n = 8), whose strongest self-reported language was not solely English (n = 10), or whose exposure phase accuracy was more than three standard deviations below the mean (n = 3; cutoff = 44/48) was removed from the statistical analysis. Certain participants were also taken out of the data set due to atypical responses in the test phase. As discussed in 2.1.3, the test phase consisted of a categorization task, where listeners heard two blocks of a 5-step *buy-pie* continuum. For each participant, the responses for both blocks were combined, and a 'difference score' was calculated by subtracting the percentage of *pie* responses at Step 5 (expected to be near 100%) by the percentage of *pie* responses at Step 1 (expected to be near 0%). Anyone whose difference score was more than three standard deviations below the mean was excluded (n = 15; cutoff = 72.71%).

The final analysis included responses from 383 subjects (231 women, 144 men, 8 non-binary; mean age = 27.68 years, sd = 4.58; self-reported ethnicity: Asian = 34, Black = 59, Latino = 11, Mixed = 55, Native Hawaiian or Pacific Islander = 1, White = 223). Given that the current experiment had two between-subjects variables with two levels each (i.e., four cells total; see 2.1.3 and 2.1.4 below for more details), there were approximately 96 listeners per cell, which doubles the sample size of a comparable perceptual adaptation study with 80% power (Cummings & Theodore, 2023).

2.1.3 Procedure

The study was conducted through a self-paced, online Qualtrics survey. After providing informed consent, participants were asked to wear headphones and to take the study in a quiet room with

no background noise. These initial instructions were followed by a sound calibration procedure (for details, refer to the Procedure portion of 2.1 in Zellou et al., 2023). Subjects then completed the main task (an exposure phase followed by a test phase) and ended the study by filling out a demographic questionnaire. Prior to each phase, participants were explicitly told that they would hear "a male speaker and a female speaker".

Both the exposure and test phases generally adhered to the procedure of Zellou et al. (2023). During exposure, listeners heard a semantically predictable sentence on each trial (e.g., "The dog had a furry paw") and were asked to identify the final keyword from one of two options: (i) the target ('paw'); (ii) a phonologically similar competitor ('pawns'). Both the target and competitor items for the critical stimuli were always real words with initial /p/, to promote the mapping from [b] to /p/ in the Female Shifted condition (see Appendix 1 for the list of competitor items).

There were 48 trials in the exposure phase (32 critical, 16 filler), which were presented in a pseudo-randomized order. The exposure stimuli were evenly divided among one male speaker and one female speaker (16 critical and eight filler each), with sentence content and talker being evenly counterbalanced. Participants were randomly assigned to either a Female Shifted or No Shift condition, which varied the production of critical items across exposure talkers (the production of filler items remained the same across conditions). In the Female Shifted condition, the female speaker produced /p/ in the critical items as [b] (e.g., "The novel is now in *brint*"), while the male speaker produced a canonical /p/ for all critical items. Both exposure speakers in the No Shift condition produced a canonical /p/ across all critical items.

Note that presenting a keyword identification task in the exposure phase, as opposed to the more commonly used lexical decision task (Kraljic & Samuel, 2007; Tamminga et al., 2020), is an important methodological choice. Lexical decision is appropriate when the critical items contain an *ambiguous* sound (e.g., 'croco?ile', where ? is between /d/ and /t/) and are thus still interpreted as real words for the most part (Kraljic & Samuel, 2007). However, the critical items in the current study have entirely *remapped* sounds (e.g., 'print' as *brint*), meaning that they would ordinarily be considered nonwords when presented in isolation (as in a typical lexical decision task). Given that perceptual adaptation is absent or reduced when critical items are interpreted as non-words (Babel et al., 2019; Norris et al., 2003), an alternative task is needed to bias listeners into perceiving bad map stimuli as real words (Charoy & Samuel, 2023; Sumner, 2011). The identification task in the current study is effective because listeners can leverage prior semantic context to deduce the identity of any shifted word (e.g., given the sentence "The novel is now in *brint*", listeners can use the prior word 'novel' to deduce that *brint* is intended as 'print'; Zellou et al., 2023). Moreover, given the focus of the current study on a gender-mediated phonetic shift, an added benefit of

using sentence-length exposure stimuli (as opposed to isolated words) is that participants can potentially hear even more acoustic cues to speaker gender (e.g., through greater exposure to suprasegmental information; Holliday, 2021).

After the exposure phase, participants completed a test phase. Listeners heard a single stimulus on each trial and were asked whether they heard *buy* or *pie*. The test phase consisted of two blocks that each presented 45 trials in a pseudo-randomized order. The test stimuli came from a 5-step *buy-pie* continuum, with each step being heard nine times per block. One block presented a novel female speaker, while the other block presented a novel male speaker. Test block order was evenly counterbalanced, such that participants either heard the novel female speaker first (Female → Male) or the novel male speaker first (Male → Female). The three female and three male speakers referenced in 2.1.1 were always evenly selected across listeners as either an exposure or test talker.

2.1.4 Analysis

Test phase responses were coded binomially as *pie* (= 1) or *buy* (= 0) and analyzed with Bayesian mixed-effects logistic regression in R (R Core Team, 2021) using the *brms* package (Bürkner, 2017) and *Stan* (Stan Development Team, 2023). No statistical analysis was conducted on the exposure phase in any experiment, given that accuracy was essentially at ceiling (99.72% across all conditions).

The model contained main effects of Step (within-subjects; scaled and centered), Speaker Gender (within-subjects; Female, Male), Block Order (between-subjects; Female → Male, Male → Female), and Exposure Condition (between-subjects; Female Shifted, No Shift). Step was treated as a numeric variable, while the other main effects were sum-coded, categorical variables. All possible interactions were included. The random effects structure consisted of by-speaker and by-listener random intercepts, as well as by-listener random slopes for Step, Speaker Gender, and their interaction. For clarity, the model structure in R syntax is shown in Equation (1).

(1) Response ~ Step * Speaker Gender * Block Order * Exposure Condition + (1 + Step * Speaker Gender | Listener) + (1 | Speaker)

Effects are interpreted as meaningful if 95% credible intervals do not contain zero (Vasishth et al., 2018). Following recent work on Bayesian mixed-effects logistic regression (Aoki & Zellou, 2024; Barreda & Silbert, 2023), the prior distributions in R syntax for the intercept, non-intercept fixed effects (b), and the standard deviation of the random intercepts (sd) were all set to: student_t(3, 0, 3).

2.2 Results

Table 4 and **Figure 1** present the aggregated test phase results and model summary statistics, respectively. There was a consistent effect of Step, which demonstrates that as continuum step increased, listeners selected *pie* more often. There were no other meaningful main effects or interactions. Notably, no interactions between Speaker Gender and Exposure Condition surfaced, implying that the exposure conditions (Female Shifted and No Shift) did not differ in *pie* responses by speaker gender.

Table 4: Summary statistics for the statistical model in Experiment 1. Meaningful effects are in bold.

	Estimate	**Est. Error**	**l-95% CI**	**u-95% CI**
Intercept	**1.49**	**0.72**	**0.04**	**2.94**
Step	**4.92**	**0.12**	**4.67**	**5.16**
Speaker Gender (Female)	0.20	0.73	–1.34	1.67
Block Order (Female → Male)	–0.14	0.12	-0.38	0.08
Exposure Condition (Female Shifted)	0.08	0.12	–0.14	0.31
Step : Speaker Gender	–0.01	0.09	–0.19	0.16
Step : Block Order	–0.14	0.11	–0.35	0.07
Speaker Gender : Block Order	0.00	0.07	–0.14	0.14
Step : Exposure Condition	0.06	0.11	–0.16	0.26
Speaker Gender : Exposure Condition	–0.02	0.07	–0.16	0.13
Block Order : Exposure Condition	0.05	0.12	–0.17	0.28
Step : Speaker Gender : Block Order	–0.09	0.07	–0.22	0.04
Step : Speaker Gender : Exposure Condition	0.02	0.07	–0.11	0.15
Step : Block Order : Exposure Condition	0.05	0.11	–0.17	0.27
Speaker Gender : Block Order : Exposure Condition	0.07	0.07	–0.07	0.21
Step : Speaker Gender : Block Order : Exposure Condition	0.04	0.07	–0.09	0.16

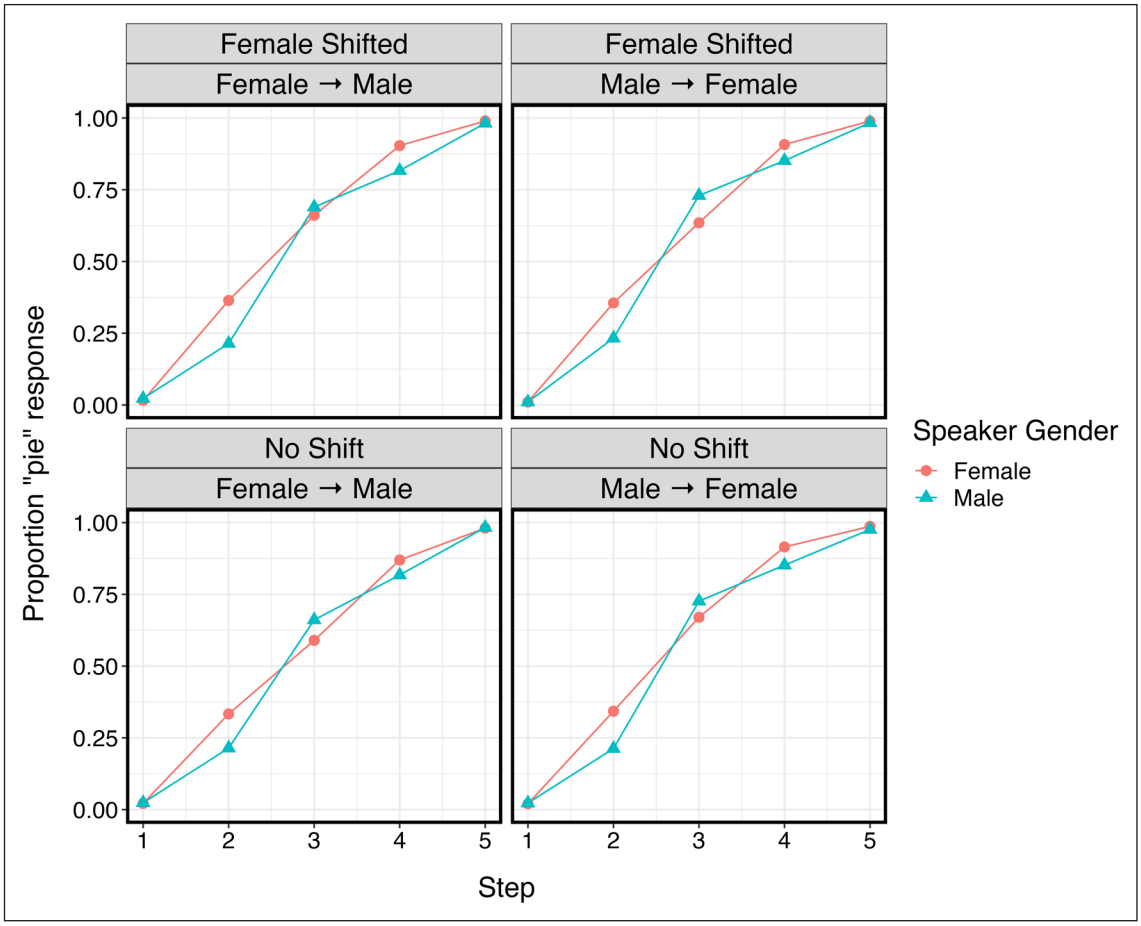

Figure 1: Aggregated results by Step, Speaker Gender (female = red, male = blue), Exposure Condition (Female Shifted, No Shift), and Test Block Order (Female → Male, Male → Female) in Experiment 1.

2.3 Experiment 1: Interim discussion

Experiment 1 placed listeners in either a critical or control condition (labelled as the Female Shifted and No Shift conditions, respectively). Both conditions exposed participants to a single male and a single female talker, but differed in whether the female talker produced a /p/ to [b] phonetic shift (the critical condition) or remained unshifted (the control condition). The male exposure talker was unshifted for all listeners. In the test phases for both the critical and control conditions, no difference in *buy-pie* categorization was observed between the novel female and novel male talkers, reflecting a lack of cross-talker generalization.

The clearest finding from Experiment 1 is that the results go against a sufficient similarity account (Xie & Myers, 2017; Xie et al., 2021). Generalization did not occur from the phonetically shifted female exposure talker to the novel female test talker, even though both were similar in terms of acoustics (higher f0) and social category (same gender). The absence of generalization

is especially notable given the wording of the instructions in the present study. Speaker gender was explicitly mentioned before both the exposure and test phases, which in theory could have biased listeners towards generalization based on shared social category membership.

The Experiment 1 results could be explained by one of two theoretical accounts. First, in accordance with a talker normalization account (Joos, 1948; Liberman & Mattingly, 1985), perhaps listeners in the critical condition overlooked social cues in the speech signal, disregarding the systematic gender covariation in the exposure phase (i.e., that the /p/ to [b] phonetic shift was only produced by the female exposure talker, not the male exposure talker). If speaker gender is ignored, then the exposure phase would be perceived as containing conflicting information, where /p/ is produced as a canonical [p] half of the time and as [b] for the other half of trials. This type of scenario has resulted in a null effect in recent work (e.g., Tzeng et al., 2021), which would be realized in the current study as no difference in test phase categorization across talkers and conditions.

Alternatively, the absence of generalization in Experiment 1 could be explained by a numerosity account (e.g., Bradlow & Bent, 2008). The /p/ to [b] shift was designed to be unattested in L1-English (see 1.3 for details), and since the production of /p/ as [b] was only heard in one exposure talker, listeners may have considered the shift to be a talker-specific trait that does not generalize to a novel talker. According to a numerosity account, generalization should only occur if listeners hear the /p/ to [b] shift in more than one talker, as it would confirm that the shift is not merely an idiosyncrasy.

Although the findings of Experiment 1 challenge a sufficient similarity account and match the predictions of talker normalization and numerosity accounts, (see **Table 1**), a null effect is not sufficient evidence to make a firm conclusion (Vasishth & Gelman, 2021). This issue is addressed in Experiment 2, which is designed to tease apart the talker normalization and numerosity accounts.

3. Experiment 2: Exposure to multiple talkers per gender

Experiment 2 has the exact same design as the previous experiment, except that participants were exposed to two female and two male talkers (instead of just one female and one male talker). Listening to *multiple* female talkers producing /p/ as [b] constitutes greater evidence that rather than being a talker-specific trait, the phonetic shift is generally associated with female talkers (i.e., there is covariation between a linguistic feature and a broader category of speakers). If a numerosity account is supported, robust cross-talker generalization should be observed, with listeners providing more *pie* responses for the female test talker than for the male talker in the critical exposure condition. A talker normalization account, meanwhile, predicts that listeners should ignore speaker gender, leading to no difference in *buy-pie* categorization between the novel test talkers.

3.1 Methods

3.1.1 Stimuli

The stimuli were exactly the same as in Experiment 1.

3.1.2 Participants

420 participants, none of whom completed Experiment 1, were recruited on Prolific and provided informed consent. The amount of compensation and the demographic filters were the same as in Experiment 1. Participants were excluded from the analysis if they either self-reported a hearing difficulty (n = 8), self-reported that their strongest language was not solely English (n = 13), self-reported being older than 35 years old (i.e., a mismatch from their official Prolific profile; n = 1), had an exposure phase accuracy more than three standard deviations below the mean (n = 3, cutoff = 44/48), or had a test phase difference score more than three standard deviations below the mean (n = 9, cutoff = 70.42%; for details on this measure, see 2.1.2). After exclusions, the Experiment 2 data set consisted of responses from 386 participants (191 women, 184 men, 11 non-binary; mean age = 28.61 years, sd = 4.43; self-reported ethnicity: Asian = 59, Black = 58, Latino = 19, Mixed = 45, Native American or Alaska Native = 3, Native Hawaiian or Pacific Islander = 2, White = 200).

3.1.3 Procedure

The procedure largely mirrored that in Experiment 1. The only difference was that, instead of presenting one male and one female speaker in the exposure phase, two male and two female exposure speakers were presented in Experiment 2 (the number and gender of the exposure talkers were again explicitly mentioned in the instructions). All four exposure talkers in the No Shift control condition produced /p/-initial words with a canonical [p]. By contrast, both female exposure talkers in the Female Shifted condition produced /p/-initial words with a canonical [b], while both male exposure speakers were unshifted. A critical point is that the relative amount of exposure to /p/-shifted words remained the same across listeners in the critical conditions of Experiment 1 and Experiment 2 – the /p/-shifted stimuli in the latter experiment were evenly distributed across two female talkers, rather than just produced by one female talker.

3.1.4 Analysis

The statistical model was identical to that in Experiment 1.

3.2 Results

Exposure phase accuracy was nearly at ceiling (99.70% across all conditions). The model summary statistics and aggregated test phase results for Experiment 2 are shown in **Table 5** and **Figure 2**, respectively.

Table 5: Summary statistics for the statistical model in Experiment 2. Meaningful effects are in bold.

	Estimate	**Est. Error**	**l-95% CI**	**u-95% CI**
Intercept	**1.29**	**0.62**	**0.04**	**2.50**
Step	**4.56**	**0.11**	**4.35**	**4.77**
Speaker Gender (Female)	0.17	0.62	–1.06	1.39
Block Order (Female → Male)	**–0.26**	**0.11**	**–0.48**	**–0.05**
Exposure Condition (Female Shifted)	–0.02	0.11	–0.25	0.19
Step : Speaker Gender	–0.13	0.09	–0.30	0.04
Step : Block Order	–0.01	0.10	–0.20	0.18
Speaker Gender : Block Order	0.13	0.08	–0.02	0.28
Step : Exposure Condition	–0.08	0.10	–0.27	0.12
Speaker Gender : Exposure Condition	0.09	0.07	–0.05	0.23
Block Order : Exposure Condition	–0.05	0.11	–0.26	0.16
Step : Speaker Gender : Block Order	–0.12	0.07	–0.25	0.01
Step : Speaker Gender : Exposure Condition	–0.03	0.07	–0.15	0.11
Step : Block Order : Exposure Condition	–0.12	0.10	–0.31	0.07
Speaker Gender : Block Order : Exposure Condition	**0.18**	**0.07**	**0.03**	**0.32**
Step : Speaker Gender : Block Order : Exposure Condition	0.08	0.07	–0.05	0.21

A meaningful main effect of Step emerged, indicating that listeners provided more *pie* responses as step number increased. There was also a consistent main effect of Block Order, such that on average, listeners who heard the female block first made fewer *pie* responses. Critically, however, the effect of Block Order was modulated by a 3-way interaction between Speaker Gender, Block Order, and Exposure Condition.

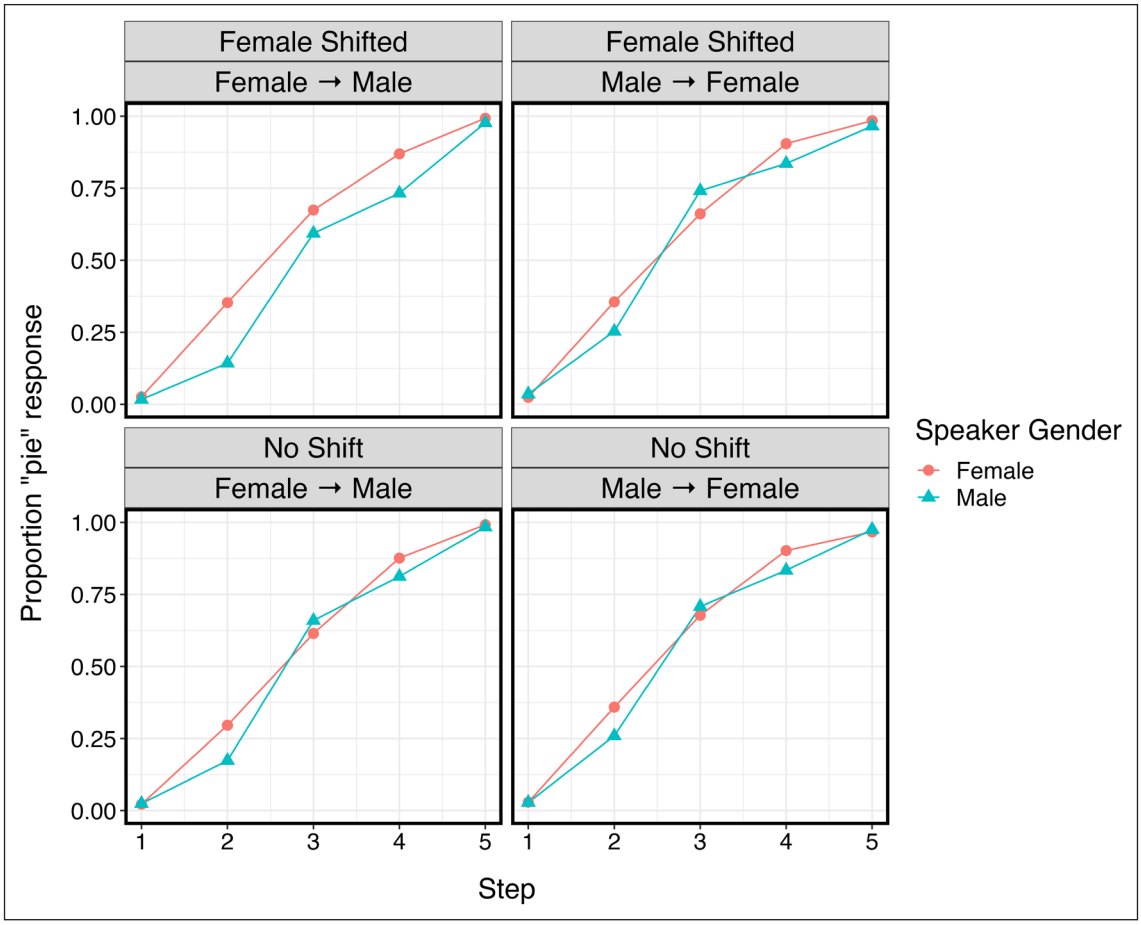

Figure 2: Aggregated results by Step, Speaker Gender (female = red, male = blue), Exposure Condition (Female Shifted, No Shift), and Test Block Order (Female → Male, Male → Female) in Experiment 2.

To examine this 3-way interaction, the *hypothesis* function from *brms* was employed (a method of inspecting interactions without running additional post-hoc models, as in a frequentist analysis; for details, see Chapter 7 of Barreda & Silbert, 2023). The hypothesis compared the 2-way interaction between Speaker Gender and Exposure Condition for each block order. A meaningful interaction with a positive coefficient was found for listeners who heard the female block first [β: 0.26, SE: 0.11, 95% HDI = (0.05, 0.47)], but not for listeners who heard the male block first [β: –0.09, SE: 0.10, 95% HDI = (–0.29, 0.11)]. In other words, relative to the No Shift control condition, participants in the Female Shifted exposure condition gave more *pie* responses for the female speaker than for the male speaker, but only if they heard the female block first (i.e., Female → Male, not Male → Female).

Looking just at the top-left graph of **Figure 2**, it becomes visually evident how listeners assigned to the Female Shifted exposure condition and Female → Male test block order provided a greater *pie* response for the female speaker relative to the male speaker. However, when this top-left cell is compared to the other three cells within **Figure 2**, the adaptation effect might initially appear to be more nuanced – the three-way interaction seems to be driven by a reduced *pie* response for the novel male speaker, rather than a greater *pie* response for the novel female speaker. This was investigated statistically through an additional hypothesis examining the 2-way interaction between Block Order and Exposure Condition for each gender. There is ultimately no clear evidence to support this impressionistic observation of **Figure 2**, as there were no meaningful interactions found for either the male test speaker [β: –0.22, SE: 0.14, 95% HDI = (–0.49, 0.05)], or the female test speaker [β: 0.13, SE: 0.12, 95% HDI = (–0.11, 0.36)]. (That said, there is a marginal interaction for the male test speaker, which accounts for why the meaningful effect of Block Order in the main model has a negative coefficient.)

3.3 Experiment 2: Interim discussion

Unlike Experiment 1, Experiment 2 found evidence for cross-talker generalization, with listeners in the Female → Male test block order providing more *pie* responses for the novel female test talker than for the novel male test talker. The only difference between the two experiments was the number of exposure talkers – Experiment 2 presented two female and two male exposure talkers, while Experiment 1 only exposed participants to one female and one male talker. It therefore seems that exposure to multiple talkers of each gender is necessary for generalization in this case, contrary to findings reported in recent work (Xie & Myers, 2017; Xie et al., 2021).

Taken together, the results of Experiments 1 and 2 provide the clearest support for a numerosity account for cross-talker generalization. Since the phonetic shift is heard in *multiple* female speakers in Experiment 2, listeners might generate a mental model that groups the female exposure speakers together (either through shared acoustic features or shared social category membership) and designate a /p/ to [b] shift as a property of female speech in the context of the experiment. The talker-general nature of this model then results in generalization of the shift to a novel female talker, who is more similar (acoustically and socially) to the female exposure speakers. Overall, the current study finds a critical role of the number of exposure talkers on generalization, bolstering numerosity accounts and challenging both talker normalization and sufficient similarity accounts.

There are two caveats of the analysis and results in Experiments 1 and 2. First, generalization was somewhat short-lived in Experiment 2. A test phase categorization shift only occurs among listeners assigned to the Female → Male block order (i.e., when the female talker was presented

first in the test phase), not among those placed in the Male → Female order (i.e., when there is a delay in presenting the novel (female) talker who is more similar to the shifted (female) exposure talkers). There are at least three different, non-mutually exclusive explanations for this order effect.

One explanation involves the particular test case used in the critical condition of the current study. As discussed in 1.3, there is no attested covariation in L1-English between a bad map stop consonant shift and speaker gender, such that female speakers produce /p/ as [b] and male speakers remain unshifted. The lack of familiarity with this type of phonetic shift may have made it more challenging to learn, resulting in more ephemeral generalization. A more robust effect might be observed if one or more aspects of the present test case are adjusted to align with attested L1-English phenomena (e.g., if the *male* speakers produced /p/ as [b] and the female speakers were unshifted; Robb et al., 2005).

Another possibility is that perhaps there was not enough exposure to result in prolonged generalization. Listeners in the critical condition of Experiment 2 heard 16 phonetically shifted stimuli from only two female talkers during a several-minute-long exposure phase. Given the novelty of the phonetic shift, greater exposure may be needed for the shift to be retained in memory. A more stable effect could be observed if listeners were presented with more stimuli[6] (Cummings & Theodore, 2023), completed several iterations of this task over multiple sessions (Xie & Kurumada, 2024), and/or heard more than two shifted talkers (Bradlow & Bent, 2008).

Finally, the transitory nature of generalized adaptation in Experiment 2 could also be explained by the structure of the exposure phase. Listeners heard a mixture of both filler and critical stimuli from two different male and two different female talkers in a randomized order, rather than in blocks. Perhaps presenting the exposure sentences in a blocked order (either by talker or by gender) would have made it easier to keep track of the speakers and to learn that only the females were phonetically shifted. This account accords with work showing detrimental effects of talker-switching on speech perception (Magnuson et al., 2021) and with the intuition that blocked training may be more helpful at the incipient stages of learning (for discussion, see Raviv et al., 2022, p. 476).[7]

---

[6] Although increasing the number of critical stimuli should facilitate adaptation to a certain point (Cummings & Theodore, 2023), *too many* critical stimuli can lead to a negative after-effect, resulting in diminished adaptation (for details, see Kleinschmidt & Jaeger, 2016, p. 683).

[7] The effect of variability on speech perception is highly complex (Quam & Creel, 2021), and in contrast to the proposal in 3.3, Tzeng et al. (2016) found that a randomized exposure stimulus order facilitated cross-talker generalization, not a blocked order. However, the experimental design of Tzeng et al. (2016) is not the same as that of the current study. The latter authors focus on a different phenomenon (adaptation towards Spanish-accented English), using a different task (sentence-transcription-in-noise) with a different method of measuring generalization (transcription accuracy of a novel test talker). Given the differences present in Tzeng et al. (2016), it is unclear whether the same results should be expected in the current study. Future work examining the effect of training structure on perceptual adaptation is warranted.

Besides the order effect in Experiment 2, a second caveat is that so far, only the aggregated results have been shown, not the results by individual test talker. As mentioned in 2.1.3, the three female and three male talkers were fully counterbalanced so that each was presented as a test talker equally often. There is a minor possibility that the results genuinely support a sufficient similarity account (rather than a numerosity account), but that the aggregated analysis is masking this finding. The following section investigates this idea in greater depth and then presents a post-hoc analysis of both experiments.

4. Post-hoc analysis of Experiments 1 and 2

A numerosity account claims that listeners must hear multiple talkers of each gender in order to generalize a gender-mediated phonetic shift to novel speakers (Bradlow & Bent, 2008). In this case, as depicted in the hypothetical set of results in **Table 6**, there should be no evidence of cross-talker generalization for any exposure-test talker pair in Experiment 1, where only one talker of each gender was presented in exposure. By contrast, all possible exposure-test talker combinations in Experiment 2 should lead to generalization.

Table 6: Hypothetical set of results supporting a numerosity account. Each cell represents a particular exposure-test talker combination, where the column labels reflect the possible exposure talker(s) and the rows stand for the possible test talker. The second, third, and fourth columns (Joanna, Ruth, Salli) refer to Experiment 1 (one exposure talker), while the fifth, sixth, and seventh columns (Joanna + Ruth, Joanna + Salli, Ruth + Salli) refer to Experiment 2 (two exposure talkers). The values within each cell refer to presence ('Yes') or absence ('No') of cross-talker generalization in this hypothetical scenario. The dashes reflect the lack of talker-specific conditions in the current study (i.e., the exposure talkers were never presented in the test phase).

	Joanna	**Ruth**	**Salli**	**Joanna + Ruth**	**Joanna + Salli**	**Ruth + Salli**
Joanna	-----	No	No	-----	-----	Yes
Ruth	No	-----	No	-----	Yes	-----
Salli	No	No	-----	Yes	-----	-----

According to a sufficient similarity account, the degree of similarity between exposure and test talkers mediates cross-talker generalization, not the number of talkers in exposure (Xie & Myers, 2017; Xie et al., 2021). Multi-talker exposure should only enhance generalization when there happens to be at least one exposure talker that is sufficiently similar to the test talker (and by the same token, if none of the exposure talkers are similar to the test talker, then multi-talker exposure is not expected to facilitate generalization). Critically, the test talker does not need to be similar to all of the exposure talkers for generalization to take place (Xie & Myers, 2017).

One possibility is that both the null effect in Experiment 1 and the generalization effect in Experiment 2 could be explained by a sufficient similarity account, not by a numerosity account. Recall that there are three female talkers (named Joanna, Ruth, and Salli; see 2.1.1 for details). Experiment 1 presents one female talker in exposure and one novel female talker at test, meaning that there are six possible exposure-test combinations (Joanna → Ruth, Joanna → Salli, Ruth → Joanna, Ruth → Salli, Salli → Joanna, Salli → Ruth). Experiment 2 exposes listeners to two female talkers in the exposure phase and one novel female talker in the test phase, resulting in three possible exposure-test combinations (Joanna + Ruth → Salli, Joanna + Salli → Ruth, Ruth + Salli → Joanna).

If two of the female talkers are similar to each other and both are different from the third female talker, then the results in Experiments 1 and 2 could be accounted for by a sufficient similarity account. This idea is visually depicted within the hypothetical set of results in **Table 7**. If Joanna and Ruth are similar to each other, but both differ from Salli, then Experiment 1 should only show generalization for two of six exposure-test combinations (Joanna → Ruth, Ruth → Joanna). However, given the same similarity relations, two of three exposure-test combinations in Experiment 2 should result in generalization (Joanna + Salli → Ruth, Ruth + Salli → Joanna), since at least one exposure talker is similar to the test talker. In an analysis that glosses over the individual exposure-test groupings, as in 2.2 and 3.2 of the current study, Experiment 2 should show a test phase categorization shift, since the majority of participants (two-thirds) hear at least one exposure talker that is similar to the test talker. By contrast, only a minority of participants in Experiment 1 (one-third) are presented with sufficiently similar exposure and test talkers, which could lead to a null aggregate effect.

Table 7: Hypothetical set of results supporting a sufficient similarity account that would still lead to the aggregated results presented in 2.2 and 3.2. The structure of the table is the same as in Table 6, such that the row labels represent the test talkers and the column labels represent the exposure talkers. 'Yes' and 'No' refer to the presence or absence of generalization, respectively.

	Joanna	**Ruth**	**Salli**	**Joanna + Ruth**	**Joanna + Salli**	**Ruth + Salli**
Joanna	-----	Yes	No	-----	-----	Yes
Ruth	Yes	-----	No	-----	Yes	-----
Salli	No	No	-----	No	-----	-----

Although a cursory acoustic analysis was already presented in 2.1.1 confirming that all of the female speakers have higher f0 than the male speakers, this is not necessarily enough evidence to confidently conclude that the female speakers are sufficiently similar to each other. Speech

contrasts are highly multidimensional (Schertz & Clare, 2020), meaning that the female speakers in the current study could be similar or different from each other in additional variables besides f0. Since there are individual differences among listeners in cue weighting (Kapnoula et al., 2017), it is difficult in practice to determine which specific acoustic cues are being leveraged in judgments of cross-talker "similarity". Therefore, the current post-hoc analysis takes a perceptual approach, assuming that two individuals must be sufficiently similar to each other if exposure to one talker leads to cross-talker generalization for the other.

4.1 Analysis and results

The aggregated data analysis in 3.2 showed evidence of cross-talker generalization in Experiment 2 (i.e., greater *pie* response for the novel female than for the novel male speaker), but only among listeners who heard the novel female block first in the test phase. Given this order effect, the post-hoc analysis data only contained participants assigned to the Female → Male block order. Furthermore, there were no meaningful interactions involving continuum step in either experiment, so Step was removed as a predictor variable from the post-hoc models.

As depicted in Equation 2, all post-hoc models contained fixed effects of Speaker Gender (within-subjects; Female, Male), Exposure Condition (between-subjects; Female Shifted, No Shift), and their interaction, by-speaker and by-listener random intercepts, and by-listener random slopes for Speaker Gender. A separate model was fitted for each of the nine possible exposure-test talker combinations across both experiments (six models for Experiment 1, three models for Experiment 2). Although this post-hoc analysis decreases statistical power, the number of subjects analyzed in each model (Experiment 1 Models: n = 33, on average; Experiment 2 Models: n = 62, on average) is still comparable to the sample sizes of prior work with 80% power (Cummings & Theodore, 2023).

(2) Response ~ Speaker Gender * Exposure Condition + (1 + Speaker Gender | Listener) + (1 | Speaker)

If cross-talker generalization has occurred, then there should be a meaningful interaction between Speaker Gender and Exposure Condition with a positive coefficient (i.e., indicating that greater *pie* response was provided for the novel female talker than the novel male talker, but only in the Female Shifted exposure condition). The credible intervals in each post-hoc model for this interaction are listed in **Table 8**. The other fixed effects are not directly relevant for the current research question and are thus omitted for clarity (but see the Data accessibility statement to examine all of the statistical models).

Meaningful interactions were only present for two of three multi-talker exposure conditions (Joanna + Salli → Ruth, Joanna + Ruth → Salli). In other words, when there was only one exposure talker in Experiment 1, none of the individual exposure-test talker pairs facilitated cross-talker generalization.

Table 8: Summary statistics for the interaction between Speaker Gender and Exposure Condition in each post-hoc model. The first six rows are from the data from Experiment 1, while the last three rows are derived from the data from Experiment 2. Meaningful effects are in bold.

Experiment	Exposure Talker	Test Talker	Estimate	Est. Error	l-95% CI	u-95% CI
Experiment 1	Joanna	Ruth	0.03	0.07	–0.11	0.17
Experiment 1	Joanna	Salli	0.02	0.06	–0.11	0.13
Experiment 1	Ruth	Joanna	–0.04	0.07	–0.19	0.11
Experiment 1	Ruth	Salli	0.06	0.05	–0.03	0.15
Experiment 1	Salli	Joanna	–0.03	0.07	–0.17	0.11
Experiment 1	Salli	Ruth	0.03	0.06	–0.10	0.16
Experiment 2	**Joanna + Salli**	**Ruth**	**0.13**	**0.04**	**0.05**	**0.22**
Experiment 2	**Joanna + Ruth**	**Salli**	**0.10**	**0.05**	**0.002**	**0.19**
Experiment 2	Ruth + Salli	Joanna	–0.06	0.05	–0.16	0.04

4.2 Post-hoc analysis: Interim discussion

The results of the post-hoc analysis are summarized in **Table 9**. Overall, clear support is provided for a numerosity account over a sufficient similarity account. No evidence of generalization was observed between any of the six female exposure-test talker combinations in Experiment 1, where only one talker was presented in the exposure phase. A sufficient similarity account would attribute this null effect to a lack of sufficient similarity between all female talkers to each other in the current study (i.e., the Joanna voice is dissimilar to both the Ruth and Salli voices, and the Ruth voice is dissimilar to the Salli voice). In theory, if the exposure phase presents two female talkers as in Experiment 2, and both are dissimilar to the test talker, then generalization should be blocked (Xie & Myers, 2017; Xie et al., 2021).

However, this prediction is not borne out. Experiment 2 found robust evidence of cross-talker generalization, at least for two of the possible exposure-test talker combinations (Joanna + Salli → Ruth, Joanna + Ruth → Salli). This suggests that, as predicted by a numerosity account (Bradlow & Bent, 2008), exposure to multiple talkers per gender is necessary to generalize an unfamiliar phonetic shift to a novel talker.

Why was no generalization observed for the Experiment 2 participants who were exposed to Ruth and Salli and tested on Joanna? One tentative possibility is that effects of acoustic similarity

are at play. Even though Joanna has a higher f0 than all three male speakers in the current study (see **Table 3**), t-tests demonstrate that the test stimuli of Joanna have a lower f0 than both the exposure stimuli of Ruth (mean difference = 22.88 Hz, $t = 23.953$, $p < 0.001$) and the exposure stimuli of Salli (mean difference = 18.37 Hz, $t = 20.585$, $p < 0.001$). The exposure stimuli of Ruth and Salli, meanwhile, are more similar to each other acoustically. Although Salli does have a consistently greater f0 than Ruth (mean difference = 4.50 Hz, $t = 5.58$, $p < 0.001$), the mean difference is reduced compared to the pairwise comparisons between Joanna/Ruth and Joanna/Salli. Perhaps Joanna was not sufficiently similar acoustically to Ruth and Salli, which blocked generalization among listeners in Experiment 2 who were exposed to Ruth and Salli and later tested on Joanna.

However, a sufficient similarity account is not entirely satisfactory. Assuming that Ruth and Salli are more similar to each other, it becomes unclear why no generalization took place in Experiment 1 for listeners who were exposed to Ruth and tested on Salli (Ruth → Salli) or for participants exposed to Salli and tested on Ruth (Salli → Ruth). Additionally, a sufficient similarity account would predict generalization to be *weaker* among listeners in Experiment 2 who were either exposed to Joanna and Salli and tested on Ruth (Joanna + Salli → Ruth) or exposed to Joanna and Ruth and tested on Salli (Joanna + Ruth → Salli). All participants in the critical conditions across Experiments 1 and 2 heard the same number of phonetically shifted stimuli (n = 16). Whereas all of these /p/ to [b] shifted tokens were produced by a single female exposure talker in Experiment 1, they were evenly distributed across two female exposure talkers in Experiment 2 (n = 8 each). If Ruth and Salli are truly the only two female speakers who share sufficient similarity, then the Experiment 2 listeners in the Joanna + Salli → Ruth and Joanna + Ruth → Salli conditions were exposed to fewer sufficiently similar tokens than the Experiment 1 listeners in the Salli → Ruth and Ruth → Salli conditions (n = 8 in Experiment 2 versus n = 16 in Experiment 1). Reduced exposure generally leads to less robust adaptation (Cummings & Theodore, 2023), so according to a sufficient similarity account, generalization should have been less strongly facilitated in the Joanna + Salli → Ruth and Joanna + Ruth → Salli conditions of Experiment 2 compared to the Salli → Ruth and Ruth → Salli conditions of Experiment 1. Yet, the opposite pattern is depicted in **Table 8** – generalization was only facilitated in the Joanna + Salli → Ruth and Joanna + Ruth → Salli conditions of Experiment 2 and blocked in the Salli → Ruth and Ruth → Salli conditions of Experiment 1. Overall, the acoustic differences between Joanna and Ruth/Salli do not appear to adequately explain the current results.

An alternative explanation could involve the properties of the test continua themselves, which are depicted in **Figure 3**. Although efforts were made to match the continua as closely as possible (see Appendix 2), **Figure 3** shows that listeners in the No Shift condition (where no phonetic shift was presented in the exposure phase) are already heavily biased towards a *pie* response for Joanna compared to Ruth and Salli (Joanna: 67.86% *pie* response; Ruth: 51.93%

pie response; Salli: 51.89% *pie* response).[8] The presence of generalization is also indicated by a greater proportion of *pie* responses, so it is possible that there was a ceiling effect – among the Experiment 2 listeners who were tested on Joanna, listeners in the Female Shifted condition could not provide a greater *pie* response than the No Shift condition, resulting in a null effect. The test phase stimuli for Ruth and Salli are not nearly as biased towards a *pie* response, so participants tested on Ruth and Salli in Experiment 2 showed meaningful generalization.

Although it is true that there are certain acoustic disparities among the female speakers (e.g., Joanna has a lower f0 than Ruth and Salli), the differences in f0 between the female and male speakers are much starker (see **Table 2**). It therefore seems more probable for listeners to group Joanna with one of the other female speakers than with the male speakers, based on (relative) acoustic and social similarity. All participants in the critical condition of Experiment 2 heard two shifted female and two unshifted male talkers in the exposure phase, and in the Joanna + Salli → Ruth and Joanna + Ruth → Salli conditions, listeners likely placed Joanna within the same mental model as Salli/Ruth. In accordance with a numerosity account, listeners attuned to the number of talkers with the /p/ to [b] shift and, given that the shift was heard in more than one exposure speaker, participants then generalized the shift to the most similar test talker (i.e., the novel female speaker).

Table 9: Actual results of the post-hoc analysis. The structure of this table is the same as Tables 6 and 7, such that the row labels represent the test talkers and the column labels represent the exposure talkers. 'Yes' and 'No' refer to the presence or absence of generalization, respectively.

	Joanna	**Ruth**	**Salli**	**Joanna + Ruth**	**Joanna + Salli**	**Ruth + Salli**
Joanna	-----	No	No	-----	-----	No
Ruth	No	-----	No	-----	Yes	-----
Salli	No	No	-----	Yes	-----	-----

[8] As discussed in Appendix 2, a norming study was conducted and efforts were made to make the continua as even as possible across speakers. These efforts were successful overall, given that there was no effect of Speaker Gender in the No Shift condition in the aggregated, group-level analysis (see **Figures 1** and **2**). However, as in Experiment 2, participants in the No Shift condition of Experiment 1 who heard Joanna in the test phase were more biased towards a *pie* response, relative to listeners who heard either Ruth or Salli at test (Joanna: 71.93% *pie* response; Ruth: 51.67% *pie* response; Salli: 50.38% *pie* response). It should be noted that the norming study had fewer participants than the main experiments (n = 120 versus n = 383 in Experiment 1 and n = 386 in Experiment 2) and a between-subjects design, with participants categorizing test phase stimuli from only one of the six speakers (i.e., ~20 listeners assigned to each speaker). The comparatively lower statistical power of the norming study could explain why the *buy-pie* continuum of Joanna seemed more evenly matched to the other female speakers in the norming study, but not in the main experiments.

Taken together, the results of the post-hoc analysis are more consistent with a numerosity account than a sufficient similarity account. A numerosity account more satisfactorily rationalizes why cross-talker generalization solely occurs with multi-talker exposure in Experiment 2 and never with single-talker exposure in Experiment 1. The absence of generalization in one of the three possible exposure-test talker combinations in Experiment 2 seems to reflect idiosyncratic variation in the test phase continua, not evidence against a numerosity account.

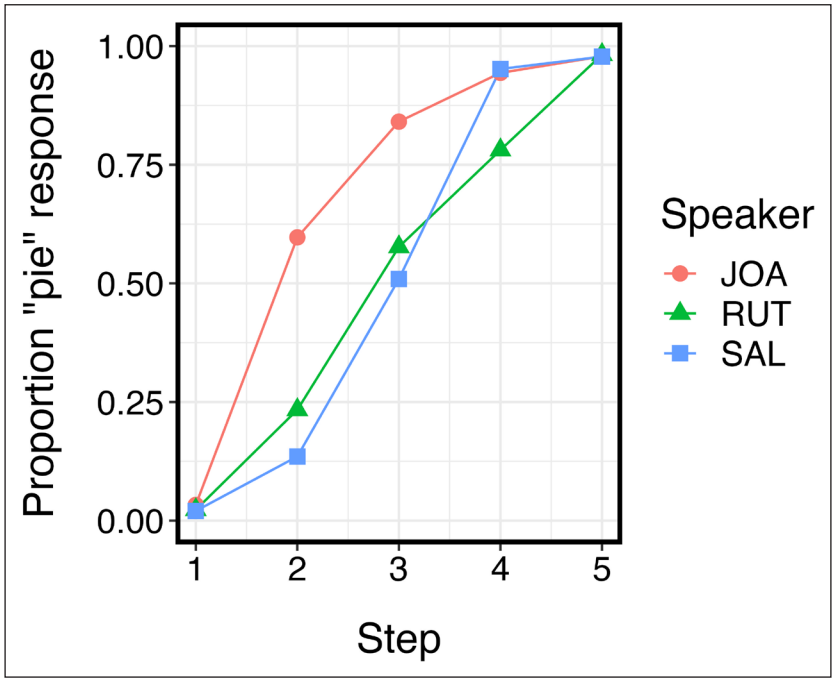

Figure 3: Aggregated results by Step and Speaker (JOA = Joanna; RUT = Ruth; SAL = Salli) in Experiment 2 for participants assigned to the No Shift condition (i.e., where no phonetic shift was presented in the exposure phase).

5. General discussion

A substantial body of work has documented extensive talker-specific and socio-indexical structure within the speech signal, with productions varying systematically based on individual talkers (Chodroff & Wilson, 2017; Newman et al., 2001) and based on social categories, such as gender, dialect, age, and many other variables (Kleinschmidt, 2019; Labov, 1966). Listeners have been consistently shown to apply their knowledge of this structure in speech perception (e.g., D'Onofrio, 2015; Niedzielski, 1999). What is not clear is how sociolinguistic perception emerges. When a listener hears a socially-conditioned variant for the first time, how do they learn that rather than being a talker-specific trait, this feature is characteristic of a broader social group that can generalize to other members of the same group?

The current study explored this question by presenting listeners with an unattested variant in L1-English (a gender-mediated /p/ to [b] phonetic shift). Experiment 1 exposed listeners to a single shifted female talker and a single unshifted male talker, while Experiment 2 exposed listeners to two shifted female talkers and two unshifted male talkers. In contrast to Experiment 1, cross-talker generalization (measured as a shift in /p/-/b/ categorization for a novel female talker) was only found in Experiment 2. Taken together, these results support a *numerosity account*. The number of exposure talkers can be critical in perceptual adaptation, and when listeners are presented with a previously unheard variant, multiple talkers per social group seem to be necessary for cross-talker generalization.

The rest of the general discussion is structured as follows. 5.1 leverages a sociolinguistic explanation to account for seeming discrepancies between the current study and prior work about how the number of exposure talkers impacts generalization. Then, 5.2 discusses a key step for future work.

5.1 Taking a sociolinguistic perspective: When is multi-talker exposure necessary for cross-talker generalization?

The present findings address the ongoing debate about whether multi-talker exposure is necessary for cross-talker generalization. In a seminal paper, Bradlow and Bent (2008) offered evidence in support of this claim: comprehension in noise of a novel Mandarin-accented English talker only improved for listeners exposed to multiple talkers of the same accent, not for listeners exposed to a single talker. In a replication of Bradlow and Bent (2008), however, Xie et al. (2021) did not observe any difference in perceptual performance between single and multi-talker exposure conditions. This later finding corroborated Xie and Myers (2017), supporting the hypothesis that the degree of acoustic similarity between the exposure and test talkers mediates generalization, not the number of talkers in exposure. Now, seemingly in contrast to Xie and Myers (2017) and Xie et al. (2021), the current study indicates that multi-talker exposure may be required for generalized adaptation in some instances.

Xie et al. (2021) account for the lack of replicability of the multi-talker exposure benefit by pointing to three methodological issues of Bradlow and Bent (2008): (i) relatively low statistical power ($n = 87$ across five between-subjects conditions); (ii) the use of different exposure talkers in the single- and multi-talker exposure conditions; (iii) the presentation of only one test talker for all participants. These factors cannot explain the current findings because a large number of listeners were recruited ($n = 383$ in Experiment 1, $n = 386$ in Experiment 2), the exposure talkers were drawn from the same pool of talkers for all conditions, and test talker identity was evenly counterbalanced. Therefore, the presence of cross-talker generalization in only Experiment 2 of the current study, when listeners heard multiple talkers per gender in exposure, cannot merely be attributed to the confounds present in Bradlow and Bent (2008).

Even though the empirical result of this work differs from Xie and Myers (2017) and Xie et al. (2021), these apparent contradictions can be reconciled by appealing to a *sociolinguistic* framework. More specifically, the types of exposure conditions that induce generalization might vary based on the amount of experience listeners have with the presented variants/accents. Both the current work and prior work might be situated along the following timeline: (i) when listeners have little or no prior experience with a variant (e.g., the current study), then multi-talker exposure is necessary for generalization; (ii) when listeners have some prior experience with a variant (e.g., work on L2-accent adaptation; Xie et al., 2017; Xie et al., 2021), then some relevant exposure is necessary for generalization, but either a single talker or multiple talkers can be heard in exposure; (iii) when listeners have extensive prior experience with a variant (e.g., work on sociolinguistic perception; Niedzielski, 1999; Strand & Johnson, 1996), then no exposure is needed for generalization.

On one end of the spectrum, the current study presented a phonetic variant that was intentionally designed to be unattested in L1-English (a "bad map" /p/ to [b] shift covarying with gender; see 1.3 for details). Given their unfamiliarity with this variant, listeners in Experiment 1 likely considered it to be a talker-specific idiosyncrasy when heard in a single female exposure talker, which blocked generalization towards a novel female talker. Generalization was only observed in Experiment 2, when participants heard the /p/ to [b] shift in multiple female exposure talkers and could thus presume that the shift was not merely a talker-specific trait. Even with multi-talker exposure, however, the generalization effect dissipates within only several minutes (see 3.2 and 3.3 for details), suggesting that listeners are reluctant to generalize when they have had little to no prior experience with a variant.

Xie and Myers (2017) and Xie et al. (2021), meanwhile, presented Mandarin-accented English to either undergraduates at a diverse American school (University of Connecticut; Xie & Myers, 2017) or L1 speakers of American English (Xie et al., 2021). Even if participants self-reported a lack of familiarity with Mandarin-accented English, many likely had at least some previous exposure to other accents or dialects, considering the prevalence of L2-accented speakers in the United States (Graddol, 2003). Participants could have therefore discerned accent-independent properties within the speech signal (e.g., slower speaking rate, difficulty producing the tense/lax vowel contrast in English; Baese-Berk et al., 2013), deducing that the acoustic features of the exposure talkers are likely part of some broader L2-accent, not entirely idiosyncratic. Compared to participants in the current experiments, listeners are much less conservative when exposed to L2-accented speech – given the appropriate conditions (e.g., sufficient acoustic similarity between the exposure and novel talkers), generalization can be facilitated with either single- or multi-talker exposure (Xie and Myers, 2017; Xie et al., 2021) and can even occur with a 12-hour delay between the exposure and test phases (Xie et al., 2017).

On the opposite end of the spectrum as the current study are experiments on sociolinguistic perception, in which listeners hear variants that they already have ample experience hearing in their everyday lives (D'Onofrio, 2015; Niedzielski, 1999; Strand & Johnson, 1996). These cases do not require any exposure phase at all for generalization to be triggered. By merely presenting a single cue to social identity, listeners can generalize their prior knowledge of sociolinguistic variation to a novel speaker. For example, participants' mental models of speaker gender and sibilant production are so robust that, without any prior exposure phase during an experiment, their categorization of tokens along an /s/-/ʃ/ continuum is altered based on whether a gender-ambiguous voice is presented with a stereotypically male or female face (Munson, 2011; Strand & Johnson, 1996).

In summary, highlighting the role of listeners' social experiences not only unifies the current study with Xie and Myers (2017) and Xie et al. (2021), but also ties perceptual adaptation to sociolinguistic perception. This explanation also adds nuance to the literature, shifting from a binary debate about *whether* multi-talker exposure is necessary for cross-talker generalization (Bradlow & Bent, 2008; Xie et al., 2017; Xie et al., 2021) to a broader discussion about *when* multi-talker exposure is necessary for cross-talker generalization.

5.2 Acoustic versus social similarity?

According to a numerosity account, listeners in the critical condition of Experiment 2 grouped the two phonetically shifted female exposure talkers together and generalized the shift towards the test talker who was the most similar (i.e., the novel female test talker, not the novel male test talker). A key question is whether this judgment of similarity is determined by acoustics or by social category membership. The current study cannot tease apart these two possibilities because female and male speakers are acoustically distinct (e.g., through differences in f0; see **Table 3**) and can also trigger differing categorical judgments about apparent gender (Hillenbrand & Clark, 2009).

Prior work on cross-talker generalization has found evidence for both types of mechanisms. As an example of a socially-mediated effect, Aoki and Zellou (2023a) exposed listeners to Mandarin-accented English and observed that generalization towards a novel Mandarin-accented English speaker was facilitated when both the exposure and novel talkers were presented with an image of an East Asian face (i.e., similarity in apparent ethnicity). Meanwhile, in the absence of visual cues, Xie and Myers (2017) suggested that generalization of Mandarin-accented English stop consonants was based primarily on acoustic similarity, considering that listeners could not identify the accent of the speakers. Both acoustic and social similarity can lead to generalization, and it is thus possible that both are motivating the effects in the current study.

Future work on cross-talker generalization could explore what factors mediate the relative weighting of acoustic versus social similarity. For instance, there might be a critical role played by the types of voices that are presented. The current work deliberately focused on gender

as a social category because it is highly salient perceptually and easily recognized by adult speakers (Hillenbrand & Clark, 2009). By contrast, Xie and Myers (2017) and Xie et al. (2021) examined cross-talker generalization of Mandarin-accented speech. While naive listeners generally demonstrate above-chance performance in identification and discrimination tasks for L2-accents, accurate categorization of a speaker's accent is often lower than for gender (Atagi & Bent, 2017). Therefore, generalization effects for gender and L2-accented speech could be inherently different, with the former recruiting the influence of social categories (since listeners can easily match exposure and test talkers by gender) and the latter relying more heavily on acoustic mechanisms (since listeners are unsure whether the exposure and test talkers actually share the same accent).

Building upon Aoki and Zellou (2023a), however, presenting socially relevant visual cues might modulate effects of acoustic similarity. Generalization between two acoustically different Mandarin-accented English talkers in Xie and Myers (2017) could be enhanced if both speakers share the same apparent ethnicity (and by the same token, a diminished generalization effect might be observed if two acoustically similar talkers do not have the same apparent ethnicity). Socially-mediated perceptual effects also do not necessarily have to be mediated by preconceived social categories – if a phonetically shifted or accented speaker has a pen in their mouth (Liu & Jaeger, 2018), for example, it might trigger generalization towards another individual who similarly has a pen in their mouth.

Another important factor could be the experimental paradigm that is employed. The instructions of the current study mentioned the number and gender of the speakers within each phase, which could have brought more attention to social category membership. Indeed, at the end of the demographic survey following the main task, a handful of listeners in the critical conditions (n = 24 across both experiments) explicitly commented on the /p/ to [b] phonetic shift (e.g., "The female voices in part one sounded like they were saying all of the 'P' words with a B instead and I thought that might've been intentional"). The phonetic shift in the current study was intended to be clearly noticeable, and the exposure phase task (keyword identification) was designed to draw attention to the phonetic shift. Stating whether stimuli are *buy* or *pie* in the test phase patently relates to the exposure phase manipulation, so listeners in the current study may have been influenced by strategic, decision-making biases based on social similarity (Xie et al., 2023). It seems plausible for listeners to take a different approach towards generalization when given a more implicit task with a less obvious phonological contrast (e.g., as in Xie and Myers (2017), who examined devoiced word-final stop consonants in Mandarin-accented speech through auditory lexical decision and cross-modal priming tasks).

Overall, as recent work has noted (Xie et al., 2023), the precise mechanisms of cross-talker generalization are still unclear, and additional work is needed to unpack effects of acoustic versus social similarity.

6. Conclusion

Recent work on L2-accented English has claimed that multi-talker exposure is not necessary for cross-talker generalization (Xie & Myers, 2017; Xie et al., 2021). However, the findings from the current study suggest that in certain cases, multi-talker exposure does appear to be necessary for generalization – listeners only apply a gender-mediated phonetic shift to novel talkers when they are previously exposed to multiple talkers per gender, not just a single talker per gender.

This seeming contradiction can be resolved by appealing to the social experiences of the listener. In particular, whereas many listeners have encountered some type of L2-accented English, the phonetic shift in the present experiments was designed to be unattested in L1-English (i.e., covariation of gender and stop consonant production, where female speakers produce /p/ as [b] and male speakers produce prototypical /p/). Multi-talker exposure might be unnecessary when exposed to more familiar types of speech and necessary when exposed to completely unfamiliar variants.

More broadly, relatively little work explicitly integrates perceptual adaptation and sociolinguistics – the present experiments begin to fill this gap, offering insights into cross-talker generalization and the emergence of sociolinguistic perception.

Appendix 1: Exposure stimuli

Table 10: The list of exposure stimuli, along with the number of words in each sentence, the target word, and the competitor item. The first 32 rows contain the critical stimuli, while the last 16 rows contain the filler stimuli.

	Sentence	N of words	Competitor
1	The wall needed a new coat of paint.	8	part
2	Instead of eating the entire cake, Joey only ate a part.	11	past
3	Martha focuses on the future without dwelling on the past.	10	pause
4	She started talking after a long pause.	7	paw
5	The dog had a furry paw.	6	pawns
6	Chess games start with sixteen pawns.	6	peace
7	The minister vowed to achieve world peace.	7	peel
8	A key ingredient in the dessert is the zest from an orange peel.	13	perk
9	Free food was a great perk.	6	pinch
10	Do not add too much salt to the stew, just add a pinch.	13	pine
11	The travelers walked through forests of oak and pine.	9	pink
12	Flamingos are animals that are pink.	6	place
13	New York City is a crowded place.	7	plain
14	The dessert was unsatisfactory as the flavor was dull and plain.	11	plan
15	To save the organization, the CEO had a plan.	9	plant
16	The seedling grew into a healthy plant.	7	plate
17	The naughty girl had no veggies on her plate.	9	plea
18	The crying man made an emotional plea.	7	plug

(Contd.)

	Sentence	N of words	Competitor
19	Charging a cell phone often requires a plug.	8	plunge
20	With her swimsuit on, Rachel took a plunge.	8	point
21	The soccer team won the match by one point.	9	pork
22	The tacos contained chicken and pork.	6	port
23	A horn announced that the ferry had left the port.	10	pose
24	At the end of the runway, the model struck a pose.	11	pouch
25	The kangaroos were nursed in the mother's pouch.	8	prank
26	Danny was a jokester who was known to love a good prank.	12	priest
27	After Sunday mass, the churchgoer chatted with the priest.	9	prince
28	The resident of the castle is a young prince.	9	print
29	The novel is now in print.	6	prize
30	In the athletic contest, the fastest runner won the grand prize.	11	pulse
31	The doctor tried to revive the girl who had no pulse.	11	purse
32	Sandra always stored her wallet in her purse.	8	paint
33	Lisa had no shovel so she could not dig.	9	sick
34	A large aquarium holds many fish.	6	switch
35	The shoes were so small that her feet would not fit.	11	wig
36	Molly received the doll as a gift.	7	dish
37	She smiled with a cheeky grin.	6	twin
38	The mother gave her child a kiss.	7	fin
39	The man who owned the mansion was rich.	8	kid
40	Gordon tried to climb over the ridge.	7	rim

(Contd.)

	Sentence	N of words	Competitor
41	He had many dishes to rinse.	6	dense
42	He would not skydive due to the risk.	8	desk
43	At the soccer game, Kayla hurt her shin.	8	thin
44	She came out to the lake for a swim.	9	swag
45	My wool socks are warm and thick.	7	fix
46	The storm reduced the large tree into a twig.	9	fig
47	The kite was lost to the wind.	7	hint
48	The genie announced he would grant one wish.	8	kick

Appendix 2: Test stimuli creation and norming

For each of the six voices used in the current experiments (female: Joanna, Ruth, Salli; male: Joey, Matthew, Stephen), the words 'buy' and 'pie' were separately typed into the Amazon Web Services console and downloaded. Similar to the exposure stimuli (see Section 2.1.1), all productions of *buy* and *pie* were generated with neural text-to-speech synthesis, converted from .mp3 to .wav files (Tomar, 2006), and set to 60 dB SPL (Boersma & Weenink, 2021). A 9-step *buy-pie* continuum was created in Praat for each voice, using the original *buy* and *pie* stimuli as endpoints (Winn, 2022). To create the most naturalistic stimuli possible, voice onset time (VOT) and f0 were covaried at each step, meaning that both VOT and f0 increased as the stimuli became more *pie*-like (Clayards, 2017). The VOT and f0 values for each speaker at each continuum step are shown in **Table 11** and **Table 12**, respectively.

Table 11: Voice onset time (ms) of each speaker at each continuum step. The second, third, and fourth columns correspond to the female speakers (Joanna, Ruth, Salli), while the last three columns correspond to the male speakers (Joey, Matthew, Stephen).

Step	Joanna	Ruth	Salli	Joey	Matthew	Stephen
1	9.81	14.88	10.66	14.65	8.34	12.10
2	14.17	18.99	15.59	17.76	13.23	14.28
3	22.20	25.68	24.21	24.50	20.52	22.39

(Contd.)

Step	Joanna	Ruth	Salli	Joey	Matthew	Stephen
4	31.37	30.51	30.78	32.12	28.77	29.26
5	35.10	38.77	39.11	32.54	35.44	33.83
6	44.49	45.92	46.78	44.65	43.27	43.84
7	52.02	54.30	56.16	52.08	51.94	52.91
8	60.97	56.75	62.72	57.38	58.55	61.42
9	67.77	68.83	71.31	64.23	66.69	68.22

Table 12: Onset f0 (Hz) of each speaker at each continuum step. The second, third, and fourth columns correspond to the female speakers (Joanna, Ruth, Salli), while the last three columns correspond to the male speakers (Joey, Matthew, Stephen).

Step	Joanna	Ruth	Salli	Joey	Matthew	Stephen
1	173.92	227.11	199.62	104.71	135.59	139.48
2	176.29	228.81	204.68	107.29	136.49	141.14
3	179.72	232.07	206.43	110.18	140.27	142.61
4	183.23	235.08	209.25	113.48	142.07	148.73
5	186.25	238.66	212.69	116.08	144.90	148.89
6	188.10	241.63	214.87	119.37	148.46	153.37
7	190.37	244.95	217.28	121.45	150.30	157.53
8	194.69	247.63	221.03	124.00	152.73	159.85
9	195.90	251.62	222.93	126.57	154.38	164.16

Using the same demographic filters as in Experiments 1 and 2 (see 2.1.2 for details), 120 listeners were recruited from Prolific and paid $0.80 (approximately $9 per hour) to norm the stimuli. None of the listeners in the norming study participated in either of the main experiments. After giving informed consent, participants completed a categorization task in which they listened to a stimulus on each trial and stated whether they heard *buy* or *pie*. Each

listener was randomly assigned to one of the six speakers and categorized each continuum step five times in a randomized order (9 steps * 5 repetitions = 45 total trials). Data from five subjects were removed who either self-reported a hearing difficulty (n = 2), self-reported being older than 35 years old (i.e., a mismatch from their official Prolific profile; n = 1) or whose first language was not solely English (n = 2). 115 participants were included in the final analysis of the norming data (53 women, 60 men, 2 non-binary; mean age = 29.27 years, sd = 4.12; self-reported ethnicity: Asian = 27, Black = 15, Latino = 3, Mixed = 11, White = 59).

The aggregated results by speaker are shown in **Figure 4** and by speaker gender in the left panel of **Figure 5**. Impressionistically, the left panel of **Figure 5** indicates that while the results for Steps 1–3 and 7–9 are evenly matched by speaker gender, Steps 4–6 induce a greater *pie* response for the female speakers than for the male speakers overall. Note that in Experiments 1 and 2, evidence for adaptation constitutes a greater *pie* response for novel female speakers compared to novel male speakers (see 1.4.2 for details) – this implies that the responses at each continuum step should ideally be matched by speaker gender in the norming phase, so that any gender difference in *pie* responses in the actual experiments can be attributed to an adaptation effect, rather than an inherent property of the stimuli.

To match the proportion of *pie* responses by gender as closely as possible, five specific steps were selected for each speaker (Joanna, Ruth, Stephen: 1, 4, 5, 6, 9; Salli: 1, 3, 4, 6, 9; Joey: 1, 3, 5, 6, 9; Matthew: 1, 4, 6, 7, 9). The aggregated 5-step continua are shown by speaker gender in the right panel of **Figure 5**. To confirm the impressionistic observation that the continua are gender-matched, a Bayesian mixed-effects logistic regression model was fitted (*pie* = 1, *buy* = 0) through *brms* (Bürkner, 2017) and Stan (Stan Development Team, 2023) in R (R Core Team, 2021) using the same priors as in the main experiments (see 2.1.4). As shown in Equation 3, the model had fixed effects of Step (within-subjects; numeric, scaled and centered), Speaker Gender (between-subjects; sum-coded; Female, Male) and their interaction, as well as by-speaker and by-listener random intercepts and by-listener random slopes for Step. The model revealed a meaningful effect of Step [β: 4.32, SE: 0.29, 95% highest density interval (HDI) = (3.79, 4.91)], indicating that *pie* responses increased alongside continuum step. Critically, there was no main effect of Speaker Gender [β: 0.05, SE: 0.54, 95% HDI = (–1.02, 1.09)] and no interaction [β: 0.01, SE: 0.22, 95% HDI = (–0.43, 0.44)]. Given the lack of gender effect, the 5-step continua were deemed to be suitable and used in the test phase across all experiments.

(3) $\text{Response} \sim \text{Step} * \text{Speaker Gender} + (1 + \text{Step} \mid \text{Listener}) + (1 \mid \text{Speaker})$

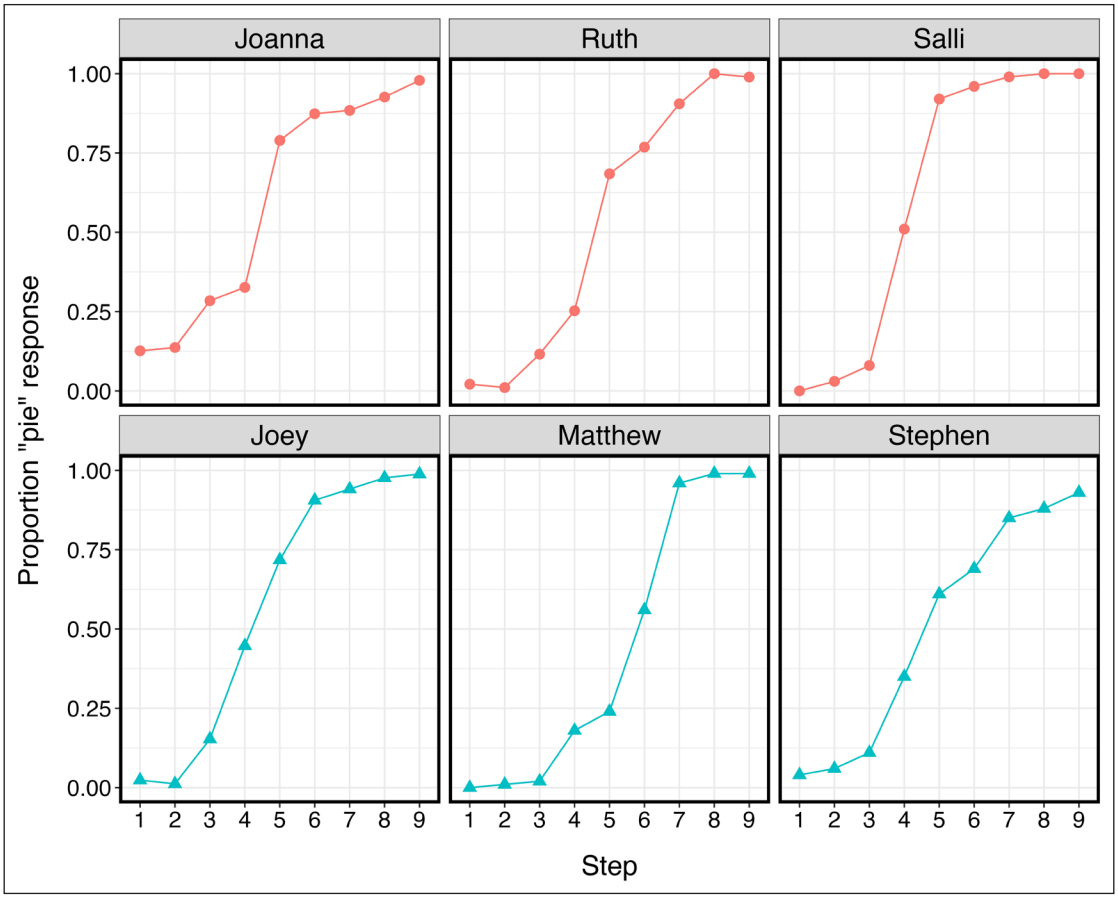

Figure 4: Aggregated results by Step and Speaker (female = top row; male = bottom row) in the norming experiment.

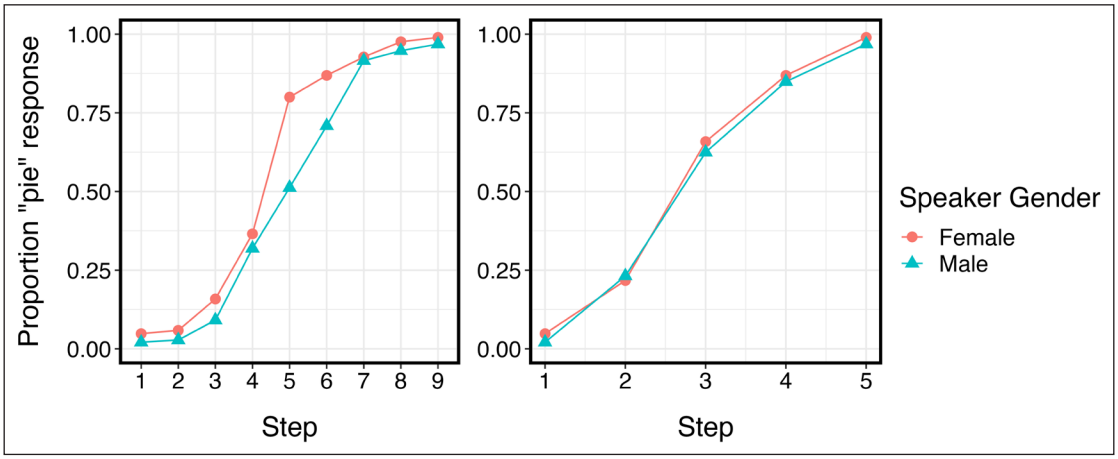

Figure 5: Aggregated results by Step and Speaker Gender (female = red; male = blue) for the original 9-step continuum (left panel) and the adjusted 5-step continuum used in Experiments 1 and 2 (right panel).

Data accessibility statement

The data and R scripts used to perform the analyses and generate the graphs are available on the Open Science Framework (OSF) at the following link: https://doi.org/10.17605/osf.io/x9jgw.

Ethics and consent

The procedures used in this study adhere to the tenets of the Declaration of Helsinki. Approval was obtained from the Institutional Review Board of the University of California, Davis (reference number: 1328085-2). The identity of all research subjects has been anonymized. All participants provided informed consent prior to entering the study.

Competing interests

The authors have no competing interests to declare.

Author contributions

NBA: Conceptualization, Data curation, Methodology, Investigation, Software, Formal analysis, Visualization, Investigation, Writing – original draft, Writing – review & editing.

GZ: Conceptualization, Methodology, Writing – review & editing, Supervision.

ORCiD IDs

NBA: https://orcid.org/0000-0002-6267-281X
GZ: https://orcid.org/0000-0001-9167-0744

References

Aoki, N., & Zellou, G. (2023c). Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 2294–2298). Guarant International.

Aoki, N. B., Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters, 2*(4), 045204. https://doi.org/10.1121/10.0010274

Aoki, N. B., & Zellou, G. (2023a). Visual information affects adaptation to novel talkers: Ethnicity-specific and -independent learning of L2-accented speech. *The Journal of the Acoustical Society of America, 154*(4), 2290–2304. https://doi.org/10.1121/10.0021289

Aoki, N. B., & Zellou, G. (2023b). When clear speech does not enhance memory: Effects of speaking style, voice naturalness, and listener age. *Proceedings of Meetings on Acoustics, 51*(1), 060002. https://doi.org/10.1121/2.0001766

Aoki, N. B., & Zellou, G. (2024). Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners. *Journal of Phonetics, 104*, 101328. https://doi.org/10.1016/j.wocn.2024.101328

Atagi, E., & Bent, T. (2017). Nonnative accent discrimination with words and sentences. *Phonetica, 74*(3), 173–191. https://doi.org/10.1159/000452956

Babel, M., McAuliffe, M., Norton, C., Senior, B., & Vaughn, C. (2019). The Goldilocks zone of perceptual learning. *Phonetica, 76*(2–3), 179–200. https://doi.org/10.1159/000494929

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America, 137*(5), 2823–2833. https://doi.org/10.1121/1.4919317

Baese-Berk, M., McLaughlin, D. J., & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass, 14*(7), e12375. https://doi.org/10.1111/lnc3.12375

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133*(3), EL174-EL180. https://doi.org/10.1121/1.4789864

Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America, 138*(3), EL223–EL228. https://doi.org/10.1121/1.4929622

Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language, 96*(2), 224–254. https://doi.org/10.1353/lan.2020.0018

Barreda, S. (2021). Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change*, *33*(1), 27–53. https://doi.org/10.1017/S0954394521000016

Barreda, S., & Predeck, K. (2024). Inaccurate but predictable: Vocal-tract length estimation and gender stereotypes in height perception. *Journal of Phonetics, 102*, 101290. https://doi.org/10.1016/j.wocn.2023.101290

Barreda, S., & Silbert, N. (2023). *Bayesian multilevel models for repeated measures data: A conceptual and practical introduction in R.* Routledge. https://doi.org/10.4324/9781003285878

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer (Version 6.1.40) [Computer program]. https://www.fon.hum.uva.nl/praat/

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Calder, J., & King, S. (2022). Whose gendered voices matter?: Race and gender in the articulation of /s/ in Bakersfield, California. *Journal of Sociolinguistics, 26*(5), 604–623. https://doi.org/10.1111/josl.12584

Campbell-Kibler, K. (2010). Sociolinguistics and perception. *Language and Linguistics Compass, 4*(6), 377–389. https://doi.org/10.1111/j.1749-818X.2010.00201.x

Carignan, C., & Zellou, G. (2023). Sociophonetics and vowel and nasality. In C. Strelluf (Ed.), *The Routledge handbook of sociolinguistics* (pp. 237–260). Routledge. https://doi.org/10.4324/9781003034636-12

Charoy, J., & Samuel, A. G. (2023). Bad maps may not always get you lost: Lexically driven perceptual recalibration for substituted phonemes. *Attention, Perception & Psychophysics, 85*, 2437–2458. https://doi.org/10.3758/s13414-023-02725-1

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001

Clayards, M. (2017). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica, 75*(1), 1–23. https://doi.org/10.1159/000448809

Cohn, M., Pycha, A., & Zellou, G. (2021). Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition, 210*, 104570. https://doi.org/10.1016/j.cognition.2020.104570

Cohn, M., & Zellou, G. (2020). Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. *Proceedings of Interspeech*, 1733–1737. https://doi.org/10.21437/Interspeech.2020-1336

Cummings, S. N., & Theodore, R. M. (2023). Hearing is believing: Lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition, 235*, 105404. https://doi.org/10.1016/j.cognition.2023.105404

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics, 19*(2), 241–256. https://doi.org/10.1111/josl.12115

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change, 1*(3), 245–267. https://doi.org/10.1017/S095439450000017X

Eckert, P. (2014). The problem with binaries: Coding for gender and sexuality. *Language and Linguistics Compass, 8*(11), 529–535. https://doi.org/10.1111/lnc3.12113

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology, 4*, 148. https://doi.org/10.3389/fpsyg.2013.00148

Feng, H., & Wang, L. (2024). Acoustic analysis of English tense and lax vowels: Comparing the production between Mandarin Chinese learners and native English speakers. *The Journal of the Acoustical Society of America, 155*(5), 3071–3089. https://doi.org/10.1121/10.0025931

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics, 15*(1), 67–83. https://doi.org/10.1016/S0095-4470(19)30538-8

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. https://doi.org/10.30658/hmc.1.5

Graddol, D. (2003). The decline of the native speaker. In G. Anderman & M. Rogers (Eds.), *Translation today: Trends and perspectives* (pp. 152–167). Multilingual Matters. https://doi.org/10.21832/9781853596179-013

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*, 1150–1166. https://doi.org/10.3758/APP.71.5.1150

Holliday, N. (2021). Prosody and sociolinguistic variation in American Englishes. *Annual Review of Linguistics, 7*, 55–68. https://doi.org/10.1146/annurev-linguistics-031220-093728

Holliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication, 8*, 1116955. https://doi.org/10.3389/fcomm.2023.1116955

Joos, M. (1948). Acoustic phonetics. *Language, 24*(2), 5–136. https://doi.org/10.2307/522229

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance, 43*(9), 1594–1611. https://doi.org/10.1037/xhp0000410

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review, 23*, 678–691. https://doi.org/10.3758/s13423-015-0943-z

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science, 10*(4), 818–834. https://doi.org/10.1111/tops.12331

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Labov, W. (1966). *The social stratification of English in New York City*. Center for Applied Linguistics.

Lai, W., & Tamminga, M. (2024). Phonetics-phonology mapping in the generalization of perceptual learning. *Journal of Phonetics, 103*, 101295. https://doi.org/10.1016/j.wocn.2024.101295

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition, 174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics, 83*, 1842–1860. https://doi.org/10.3758/s13414-020-02203-y

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543–562. https://doi.org/10.1080/03640210802035357

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics, 36*(2), 308–317. https://doi.org/10.1016/j.wocn.2007.06.003

Munson, B. (2011). The influence of actual and imputed talker gender of fricative perception, revisited (L). *The Journal of the Acoustical Society of America, 130*(5), 2631–2634. https://doi.org/10.1121/1.3641410

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America, 109*(3), 1181–1196. https://doi.org/10.1121/1.1348009

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C. (2005). Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 390–413). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757024.ch16

Pycha, A., & Zellou, G. (2024). The influence of accent and device usage on perceived credibility during interactions with voice-AI assistants. *Frontiers in Computer Science, 6*, 1411414. https://doi.org/10.3389/fcomp.2024.1411414

Quam, C., & Creel, S. C. (2021). Impacts of acoustic-phonetic variability on perceptual development for spoken language: A review. *WIREs Cognitive Science, 12*(5), e1558. https://doi.org/10.1002/wcs.1558

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Science, 26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Robb, M., Gilbert, H., & Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatrica et Logopaedica, 57*(3), 125–133. https://doi.org/10.1159/000084133

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*, 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science, 11*(2), e1521. https://doi.org/10.1002/wcs.1521

ShareAmerica. (2023, December 14). *The United States is rich in languages.* https://share.america.gov/united-states-is-rich-in-languages/#:~:text=In%20the%20U.S.%2C%20the%20number,according%20to%20the%20Census%20Bureau.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306–3316. https://doi.org/10.1121/1.3101452

Solé, M.-J. (2018). Articulatory adjustments in initial voiced stops in Spanish, French and English. *Journal of Phonetics, 66*, 217–241. https://doi.org/10.1016/j.wocn.2017.10.002

Stan Development Team. (2023). *Stan modeling language users guide and reference manual, Version*. Available online at: https://mc-stan.org (accessed January 3, 2024).

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference* (pp. 14–26). De Gruyter Mouton. https://doi.org/10.1515/9783110821895-003

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition, 119*(1), 131–136. https://doi.org/10.1016/j.cognition.2010.10.018

Swartz, B. L. (1992). Gender difference in Voice Onset Time. *Perceptual and Motor Skills, 75*(3), 983–992. https://doi.org/10.2466/pms.1992.75.3.983

Tamminga, M., Wilder, R., Lai, W., & Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology, 5*, 90–122. https://doi.org/10.2218/pihph.5.2020.4439

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal, 2006*(146), 1–10. https://dl.acm.org/doi/abs/10.5555/1134782.1134792

Tzeng, C. Y., Alexander, J. E. D., Sidaras, S. K., & Nygaard, L. C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance, 42*(11), 1793–1805. https://doi.org/10.1037/xhp0000260

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review, 28*, 1003–1014. https://doi.org/10.3758/s13423-020-01840-6

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics, 59*(5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics, 71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vonessen, J., Aoki, N. B., Cohn, M., & Zellou, G. (2024). Comparing perception of L1 and L2 English by human listeners and machines: Effect of interlocutor adaptations. *The Journal of the Acoustical Society of America, 155*(5), 3060–3070. https://doi.org/10.1121/10.0025930

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology, 2*(1), 219–237. https://doi.org/10.1515/labphon.2011.007

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics.* https://doi.org/10.1093/acrefore/9780199384655.013.95

Winn, M. (2022). Make Voice Onset Time (VOT)/F0 continuum [Praat script]. http://www.mattwinn.com/praat/Make_VOT_Continuum_v33.txt

Wolfram, W., & Schilling, N. (2015). *American English: Dialects and variation*. Wiley-Blackwell.

Xie, X., Earle, F. S., & Myers, E. B. (2017). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210. https://doi.org/10.1080/23273798.2017.1369551

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex, 166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology, 150*(11), e22–e56. https://doi.org/10.1037/xge0001039

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics, 41*(5), 369–378. https://doi.org/10.1016/j.wocn.2013.06.003

Xie, X., & Kurumada, C. (2024). From first encounters to longitudinal exposure: A repeated exposure-test paradigm for monitoring speech adaptation. *Frontiers in Psychology, 15*, 1383904. https://doi.org/10.3389/fpsyg.2024.1383904

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97*, 30–46. https://doi.org/10.1016/j.jml.2017.07.005

Yu, A. C. L. (2010). Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLOS One, 5*(8), e11950. https://doi.org/10.1371/journal.pone.0011950

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics, 5*, 131–150. https://doi.org/10.1146/annurev-linguistics-011516-033815

Zellou, G., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience, 36*(10), 1298–1312. https://doi.org/10.1080/23273798.2021.1931372

Zellou, G., Cohn, M., & Pycha, A. (2023). Listener beliefs and perceptual learning. *Language, 99*(4), 692–725. https://doi.org/10.1353/lan.2023.a914191

Zellou, G., & Tamminga, M. (2014). Nasal coarticulation changes over time in Philadelphia English. *Journal of Phonetics, 47*, 18–35. https://doi.org/10.1016/j.wocn.2014.09.002