Journal of Phonetics 103 (2024) 101295

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Phonetics–phonology mapping in the generalization of perceptual learning

Wei Lai [a,*], Meredith Tamminga [b]

[a] Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Pl, Nashville, TN 37203, USA
[b] Department of Linguistics, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 1 November 2022
Received in revised form 29 December 2023
Accepted 3 January 2024

Keywords:
Perceptual learning
Generalization
Phonetics-to-phonology mapping

ABSTRACT

Previous studies on whether perceptual learning generalizes across multiple speakers have produced inconsistent results between generalization and speaker-specificity. A prior proposal is that the critical phonemes produced by two different speakers need to be phonetically similar for perceptual learning to generalize. To test this account, we investigated the perceptual generalization of sibilants across two pairs of speakers. In both cases, sibilants of the same male speaker were manipulated to induce either an /s/-favoring perceptual bias or an /ʃ/-favoring one in a training phase. We then examined whether the perceptual biases would generalize in a test phase to the /s/-/ʃ/ continua of two different female speakers, one resembling the training speaker and the other differing from the training speaker in the spectral frequency distributions of their sibilants. We found that generalization of perceptual learning occurred in both /s/-favoring and /ʃ/-favoring conditions between the speaker pair with similar sibilant productions. For the speaker pair with different sibilants, we found perceptual generalization in the /s/-favoring training condition but not in the /ʃ/-favoring condition, which is not predicted by the phonetic similarity account. To explain these unexpected results, we offer a novel phonetics–phonology mismatch account as a refinement of our understanding of when and why perceptual generalization might be blocked. The results shed light on the constant influence of the mapping between phonetics and phonology during the learning and generalization of phonetic variability.

1. Introduction

Speaker idiosyncrasy is an important dimension of variability in speech. Listeners are known to adjust their perceptual expectations efficiently to be better aligned with the speech production of specific speakers through a process of *perceptual learning* (Norris, McQueen, & Cutler, 2003). Listeners sometimes generalize what they have learned about one speaker's acoustic properties to their perception of a different speaker's speech (Kraljic & Samuel, 2006; Kraljic & Samuel, 2007; Reinisch & Holt, 2014; Xie et al., 2018). At other times, however, listeners are found to limit their perceptually-learned expectations to the speaker that triggered the learning (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), a phenomenon sometimes referred to as **speaker-specificity**. The question of when and why listeners sometimes do and sometimes do not generalize across speakers has not been resolved.

The discrepancy between generalization and specificity of perceptual learning is particularly an issue with fricatives (Kraljic & Samuel, 2005; Kraljic & Samuel, 2006). In this paper, we demonstrate how different generalization outcomes for fricatives may be related to both their phonetic properties and their relative locations in different speakers' phonological spaces. An existing account that has already been proposed to address the generalization question from a phonetic-based perspective is what we will term the **phonetic similarity account** (Kraljic & Samuel, 2005). According to this account, the perceptual learning of fricatives generalizes only if the fricatives in question are phonetically similar across the relevant talkers, as indexed by overlap in their phonetic distributions. However, this account was proposed as a post hoc explanation in previous literature and leaves open certain questions about what more precise predictions the account would make. One question that remains unclear is whether the generalization outcome between any given pair of speakers would differ based on the direction of the experimental manipulation.

* Corresponding author.
E-mail addresses: wei.lai@vanderbilt.edu (W. Lai), tamminga@ling.upenn.edu (M. Tamminga).
URLs: https://weilaiphonetics.github.io/home/ (W. Lai), https://www.meredithtamminga.com/ (M. Tamminga).

https://doi.org/10.1016/j.wocn.2024.101295
0095-4470/

Normally, in perceptual learning experiments, instances of the critical phonemes could be manipulated in different ways to induce perceptual biases toward different directions. Would the outcome of perceptual generalization vary with specific phonetic instances presented within different experimental conditions, or would it hold consistent for the given speakers regardless of specific stimulus manipulation?

In this study, we set out to test the speaker specificity account against the phonetic similarity account to understand the constraints that govern the generalization of perceptual learning. We investigate the generalization of perceptual learning of sibilants across two pairs of speakers who have either similar or different phonetic properties for the sibilants /s/ and /ʃ/; in addition, we unpack how the direction of the experimental manipulation may give rise to different relationships between the training data and the test data in terms of phonetic similarity. Looking ahead, we will show that the predictions of neither the speaker specificity nor phonetic similarity accounts are fully supported by the results of this study. We find robust generalization from training to test in three of the four critical conditions, which is not consistent with the speaker specificity account. While the phonetic similarity account correctly predicts generalization across speakers with similar productions of the critical phonemes, it fails to account for the direction-specific perceptual generalization we find between speakers who have different productions of the critical phonemes.

As a possible explanation for our unexpected pattern of generalization results, we will offer an alternative account that we call the **phonetics–phonology mismatch** account, to explain why, in certain conditions, perceptual learning may not generalize across speakers. While the details of the proposal are easiest to understand with reference to a concrete set of stimuli, the essence of the account is that listeners in a perceptual learning experiment may learn *both* a bias toward identifying a particular phonological category and a new phonetic boundary between two phonemes. We suggest that perceptual learning will generalize only when listeners can bring both of those aspects of what they have learned from the training voice into their categorization of the test stimuli; when they come into conflict, the generalization will be blocked.

2. Background

2.1. Generalization and speaker specificity in perceptual learning

The term “perceptual learning” covers two broad lines of research in the area of speech science (Samuel & Kraljic, 2009). One line focuses on the general phenomenon that exposure to certain types of stimuli that the listeners are unfamiliar with (e.g., nonnative, accented, or degraded speech) leads to improvement in listeners’ perception of speech of that type (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004). Here we are primarily concerned with the second, narrower, line, which examines how exposure to non-canonical instances of a particular speech sound can induce listeners to change the previous mapping between atypical phonetic instances and phonological categories, as indicated by a shift of categorization boundary on the phonetic continuum (Norris et al., 2003; Kraljic & Samuel, 2005). The basic procedure is to present listeners with phonetically ambiguous stimuli with contextual information to disambiguate the intended stimulus categories. Perceptual learning is then measured by a shift in the categorization boundary between the two phonemes such that the phonetic space assigned to the contextually favored phoneme is expanded. For example, if we use contextual information to make listeners believe that a non-typical /s/-like sound is just an /s/, they may expand the range of acoustic inputs they accept as /s/ to account for the new sound. The observed category boundary shifts therefore provide a quantifiable indication of the amount of perceptual adjustment as a function of experience.

In real-world perceptual learning, listeners do not cope with only one kind of speech variability at a time, and linguistic variants do not come from only one speaker or group of speakers. As more of these factors come into play, listeners unavoidably need to make decisions about whether to extend the outcomes of perceptual learning from one situation to a different situation. Previous studies have found evidence that perceptual learning may generalize to the perception of different speakers (e.g., Kraljic & Samuel, 2005; Kraljic & Samuel, 2006; Kraljic & Samuel, 2007; Reinisch & Holt, 2014; Xie et al., 2018) and different phonemes with a similar contrast (e.g., Kraljic & Samuel, 2006; Weatherholtz, 2015; Durvasula & Nelson, 2018; Chodroff & Wilson, 2020). These findings suggest that it is possible for perceptual learning to generalize. At the same time, previous studies also suggest that listeners do not generalize indiscriminately. For example, listeners do not generalize an ambiguous pronunciation if it can be attributed to an incidental external source, such as a pen in the speaker’s mouth (Kraljic & Samuel, 2011). Listeners are also less likely to generalize what they have learned about the speech of speakers whom they perceive to be linguistically unreliable or unrepresentative (such as nonnative speakers) (e.g., Lev-Ari & Peperkamp, 2014). The scope of perceptual learning generalization in different situations may reflect listeners’ complicated knowledge about structure in real-world sociolinguistic variation (Kleinschmidt, 2019), in ways that still need to be better delineated.

The current paper focuses on how broadly perceptual learning generalizes across *speakers*. The generalization of perceptual learning across speakers has attracted a substantial body of research. Many studies have reported that exposure to a single speaker’s speech is sufficient to make listeners apply their knowledge of how that speaker sounds to other speakers (Kraljic & Samuel, 2005; Kraljic & Samuel, 2006; Kraljic & Samuel, 2007; Reinisch & Holt, 2014). Studies have also examined the effect of training with multiple speakers and found that exposure to multiple talkers with the same kind of pronunciation characteristics can promote the generalization of perceptual learning to other talkers (e.g., Bradlow & Bent, 2008; Xie, Liu, & Jaeger, 2021). Nonetheless, cases where perceptual learning fails to generalize across speakers are not uncommon (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2005), which motivated the introduction of a **speaker specificity** account. This account suggests that listeners only apply the acquired perceptual adjustment when listening to the particular speaker who caused it.

One of the first studies taken to provide evidence for speaker specificity in perceptual learning is Eisner and McQueen (2005). They found that perceptual learning of a

fricative boundary did not arise when listeners are trained on stimuli from a female voice but tested on stimuli from a male voice. Another result showing absence of perceptual generalization comes from Kraljic and Samuel (2005). This study added an "unlearning" phase in between the training and test phases, in which listeners sometimes heard additional spoken input that either contains no cases of the critical phonemes or contains non-ambiguous instances of the critical phonemes as a form of corrected input. Kraljic and Samuel (2005) found that only when the unlearning input was presented in the same talker's voice did the perceptual learning effect become attenuated. The learning was unaffected by an unlearning phase with speech of a different talker (of a different gender).

Despite the apparent discrepancy between generalization and speaker-specificity in perceptual learning, most previous studies probing this question (including those just discussed) did not observe all-or-nothing evidence for generalization or speaker specificity. Instead, what they actually reported was a mix of outcomes, reflecting high *variability* in whether perceptual learning generalizes under different manipulations of the involved stimuli. Often, these studies show that perceptual learning fails to generalize across some speakers under some circumstances, but that with small adjustments, generalization can be triggered with almost the same set of stimuli (e.g., Eisner & McQueen, 2005; Reinisch & Holt, 2014). These findings suggest that the constraints on generalization may be finer-grained than simply whether the training and test data do or do not come from the same speaker.

2.2. A phonetic similarity account of the perceptual generalization of fricatives

To reconcile varying findings of generalization and specificity, Kraljic and Samuel (2005) put forward a proposal that we refer to as the **phonetic similarity** account. According to this account, perceptual learning generalizes across different speakers if and only if the training and test stimuli have sufficient overlap in the acoustic space.[1] Apparent speaker specificity, then, may be induced not by different speaker identities *per se*, but rather by phonetic dissimilarity between different speakers and thus between training and test items.

As noted above, the evidence for the phonetic similarity account mostly comes from studies that broadly find apparent speaker specificity: in some cases, these studies also acknowledge that if the critical phonemes of different speakers are acoustically similar enough, listeners do generalize across speakers after all (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). To continue with the example of Eisner and McQueen (2005), they showed that listeners generalize the perceptual learning of a female speaker's pronunciation of /s-ʃ/ to an /ɛs-ɛʃ/ continuum when the vowel /ɛ/ is spoken by either a male or a female novel speaker, as long as the fricatives were from the original talker's speech (Experiment 1 and 2). When the continuum was created entirely from the speech of a novel talker, there was no perceptual learning (Experiment 3), unless the novel talker's fricatives had been spliced into the original talker's speech during exposure (Experiment 4). Although this result is sometimes taken as evidence for speaker specificity of perceptual learning, it might alternatively be understood to highlight the specificity of the productions of the target phonemes, rather than speaker identity as indexed by voice.

Kraljic and Samuel (2005) articulated the possibility of phonetic constraints on the generalization of fricative perceptual learning in a more specific way, based on their findings of a cross-voice asymmetry in perceptual generalization. They found that training with a female speaker's fricatives transferred to a male voice successfully, but training with the male speaker's fricatives did not transfer to the female voice. Based on acoustic analyses of the fricatives' spectral properties, they suggested that this result was conditioned on acoustic overlap between the two sets of fricatives in training and in test, associated with two different speakers: the female speaker's training fricatives had their spectral energy distributed within the frequency range of the male speaker's test fricatives' spectral energy, whereas the distribution of the male training fricatives' spectral energy was distinct from those of the female speaker's test fricatives in the frequency space. Kraljic and Samuel (2005) thus proposed that listeners track the acoustic properties of each speaker's fricatives and generalize only when there is sufficient acoustic overlap between the distributions.

2.3. Unresolved questions in the generalization of perceptual learning

As the preceding subsections should make clear, we still lack a full understanding of what constraints govern the generalization of perceptual learning: when does perceptual learning generalize or not generalize, and why? The discrepancies and variable outcomes in perceptual generalization across different studies have made it difficult to draw firm conclusions, so more work is still needed to achieve a better understanding of the constraints on perceptual generalization. While many questions remain, here we outline the particular unresolved questions that the perceptual learning study reported here was designed to address.

First, we aim to provide evidence directly contrasting the speaker specificity and phonetic similarity accounts. Teasing apart these questions has been difficult in studies comparing within-speaker to cross-speaker generalization, since a same-speaker condition will of course have greater acoustic similarity between training and test items than a different-speaker condition. This is especially true when the critical phonemes are fricatives and the speakers are of different genders, because of the robust covariation between speaker gender and the acoustic properties of fricatives (e.g., Jongman, Wayland, & Wong, 2000). For both anatomical and social-indexical reasons, the spectral energy of female speakers' fricatives is usually distributed at a higher frequency region than those of male speakers' fricatives, resulting in higher spectral peak locations and spectral mean. This co-variation is so robust that listeners apply different phonetics–phonology mapping norms in sibilant identification depending on the sex of the speaker (Strand & Johnson, 1996). Strand and Johnson (1996) showed that listeners reported a shift of the perception boundary between /s/ and /ʃ/ as a function of the voice gender on the remaining vocalization of the syllable: listeners shift their perception boundary towards the /s/-sounding end when the following vowel is produced by a female-

[1] The question of how much overlap would be sufficient on this account has not been addressed.

sounding voice and towards the /ʃ/-sounding when the vowel is produced by a male-sounding voice. Indeed, in previous perceptual learning studies, failures of perceptual generalization were mostly observed with fricatives across speakers of different genders (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Kraljic & Samuel, 2007; Reinisch, Wozny, Mitterer, & Holt, 2014; Tamminga, Wilder, Lai, & Wade, 2020), whereas such generalization failures were less frequently observed either with fricatives from speakers of the same gender (see Reinisch et al., 2014; Tamminga et al., 2020, for comparison), or with other types of phonemes (such as stops) from speakers of different genders (see Kraljic & Samuel, 2007, for comparison).

To understand whether perceptual learners care about *who* produced the training items or *what* those items sounded like, we need to take this kind of covariation into account. In this study, we do this by comparing participant's perception of two *different* female test speakers after training on the *same* male speaker. In both cases the test stimuli come from a different speaker of a different gender than the training stimuli, but the two test speakers differ in how phonetically similar their fricatives are to the training speaker. This allows us to zero in on the relative phonetic properties of the training and test stimuli while holding constant the fact that these stimuli sets come from distinct talkers. On the speaker specificity account, we should not observe the generalization of perceptual learning to either test talker. On the phonetic similarity account, we should observe generalization of perceptual learning to the test talker with fricatives that are similar to those of the training talker, and weak or no generalization to the test talker with dissimilar fricatives from the training talker.

The second unresolved issue we address is a slightly less obvious one, but still directly connected to the speaker specificity and phonetic similarity accounts. This is the question of whether the generalization outcome between any given pair of speakers could differ based on the direction of the experimental manipulation. By "direction", we refer to whether an experiment about the /s/-/ʃ/ boundary is trying to teach listeners that a manipulated ambiguous sound is an /s/ in contrast to a normal /ʃ/, or that a manipulated ambiguous sound is an /ʃ/ in contrast to a normal /s/. The phonetic similarity account highlights the possibility that generalization of perceptual learning may be contingent on the relationship between the training data and the test data in acoustic space. At first glance we might expect this similarity relationship to be stable for pairs of speakers—that is, we might think that for any given fricative, two speakers either do or do not have a similar realization of that fricative. But a perhaps-underappreciated point is that in the context of perceptual learning experiments with manipulated stimuli, the question of whether listeners hear similar or dissimilar fricatives from the training and test speakers is likely to have different answers depending on the direction of the acoustic manipulation. In creating /ʃ/-favoring training stimuli, we take the naturally-lower-frequency sibilant /ʃ/ and raise it, so that listeners hear both the naturally-high /s/ and an artificially-raised /ʃ/. The result is that the range of sibilant values along the entire /s/-/ʃ/ continuum is higher than the speaker's original range. Conversely, in creating /s/-favoring training stimuli, we take the naturally-higher-frequency sibilant /s/ and lower it, so that listeners hear both the naturally-low /ʃ/ and artificially-lowered /s/. The result is that the range of sibilant values is lower than the speaker's original range. These major shifts to the overall phonetic distributions of the critical stimuli that the listeners hear raise the possibility that generalization outcomes may be different for different experimental manipulation directions.

The reason this point is of interest is that it provides another way to contrast the speaker specificity account with the phonetic similarity account. On the speaker specificity account, we expect the predicted lack of generalization to hold consistently for any given pair of speakers, regardless of what direction the stimuli were manipulated in. But on the phonetic similarity account, we expect that outcomes may vary within a pair of speakers, in a way that reflects how the stimulus manipulation direction makes the training and test items more or less phonetically similar.

Testing these predictions does, however, require us to be able to compare the magnitude of perceptual learning in the two directions. In many previous perceptual learning studies, perceptual learning has been indexed by the difference between categorization results induced by learning conditions with opposite manipulation directions. For example, the ambiguous stimuli might be more likely to be categorized as /s/ in an /s/-favoring condition than it would be in an /ʃ/-favoring condition (e.g., Kraljic & Samuel, 2006; Kraljic & Samuel, 2007). But this approach does not reveal whether the perceptual shift has been induced by training in both directions or just one, since it does not provide an independent measure of perceptual learning within each condition. It would seem simple to add a within-subject experimental baseline condition that captures categorization before any learning takes place. However, previous studies have found that running a within-subject baseline condition before exposure to training may block any learning effect that happens later (Kraljic, Samuel, & Brennan, 2008).

Our solution will be to follow several more recent studies that use a between-subjects baseline produced by different listeners than the ones doing the perceptual learning. These studies not only provide a methodological model, but also give us further reason to believe that comparing different manipulation directions may be fruitful. In a study on the perceptual learning of /s/-/f/, Zhang and Samuel (2014) used a baseline identification function generated by a group of listeners who had not been exposed to any words with ambiguous /f/ or /s/ segments. In a comparison with this baseline, they found that exposure to words with ambiguous /s/ produced a significant perceptual shift, whereas exposure to words with ambiguous /f/ did not. Another study that adopted a between-subject baseline condition is Drozdova, van Hout, and Scharenborg (2016), who similarly found asymmetric learning effects in different training directions for an /r/-/l/ contrast.

By adding a between-subject baseline condition in our study, we will be able to compare whether generalization patterns are consistent for particular speaker pairs (per the speaker specificity account), or contingent on the phonetic properties of the critical stimuli in each condition (per the phonetic similarity account). Spelling out the concrete predictions of the phonetic similarity account for the four relevant conditions (2 directions $\times$ 2 test speakers) in our experiment will be easier to do with reference to the acoustic properties of

the experimental stimuli themselves, so we revisit this point in Section 4.1 after introducing the stimulus creation methods.

3. Method

In this experiment, listeners are trained on the speech of a male speaker and then tested on the speech of one of two female speakers. The two female speakers were chosen from a group of four female voices. They were selected in such a way that one of them has sibilants with spectral energy at higher frequencies, resulting in low similarity to the male speaker's sibilants, while the other has sibilants with spectral energy at lower frequencies, showing high similarity to the male speaker's sibilants. For each of the two tested female speakers, the experiment includes three conditions: a baseline categorization condition and two learning conditions. Participants in the baseline condition performed only a phonemic categorization task with the female speaker's /s/-/ʃ/ continuum. Participants in the learning conditions first were exposed to the male speaker's speech in a training phase, and then completed the phoneme categorization task on the female speaker's /s/-/ʃ/ continuum in a test phase. Stimuli in the training phase were manipulated to lexically favor the perception of either /s/ or /ʃ/ for ambiguous sibilant tokens, depending on the condition. We compare the categorization results of the two learning conditions with results in the baseline condition within each test speaker. If the results of the training conditions significantly deviate from the baseline categorization towards the intended direction, then it suggests that perceptual learning has generalized across the male training speaker and the female test speaker.

3.1. Recording

261 spoken words were recorded from a male speaker and two female speakers, who were undergraduate students at the University of Pennsylvania and speak relatively unmarked varieties of American English. Each of them produced 34 words that contain /s/ or /ʃ/ word medially, 14 words with initial /s/ or /ʃ/ that form 7 minimal pairs, and 33 words that do not contain /s/ or /ʃ/. All the spoken words were recorded in a soundproofed recording booth, with a Yeti microphone at a sampling rate of 44.1 kHz.

3.2. Acoustic measures

Spectral moments analysis (Forrest, Weismer, Milenkovic, & Dougall, 1988) is a widely used approach to describe the energy distribution of fricative spectra (e.g., Jongman et al., 2000; Nirgianaki, 2014; Wikse Barrow, Włodarczak, Thörn, & Heldner, 2022). The first four spectral moments (M1-M4) are calculated to describe the spectral center of gravity, standard deviation, skewness and kurtosis of the fricative. For the 34 words with word-medial /s/ or /ʃ/ for all three speakers, we used Włodarczak (2022)'s Praat script to extract the four spectral moments from a 20 ms Hann window centered around the fricative midpoint. Among the four spectral moments, M1 has been shown to be inversely related to the size of front resonating cavity, such that fricatives with more anterior place of articulation, such as /s/, have higher M1 than more posterior fricatives, such as /ʃ/ (Jongman et al., 2000; Nissen & Fox, 2005; Tjaden & Turner, 1997; Shadle & Mair, 1996). Gender differences have been reported for all spectral moments in voiceless fricatives, although primarily M1, such that women produce /s/ with higher M1 than men (e.g., Jongman et al., 2000; Maniwa, Jongman, & Wade, 2009). Given the importance of the first spectral moment in indexing both the /s/-/ʃ/ contrast and speaker gender, we will focus on M1: center of gravity (COG) as the major acoustic measurement in the remainder of this paper. Other spectral moment values as well as the raw and normalized duration of /s/ and /ʃ/ produced by the three speakers can be found in Appendix A Table 2.

The COG values of the 34 words produced by each speaker (in Hz) were plotted in Fig. 1. The figure shows that the three speakers differ from each other only slightly in the COG of the 17 /ʃ/ sounds, with the two female speakers having higher COG for /ʃ/ than the male speaker by around 250–500 Hz. By contrast, the three speakers differ substantially in their /s/ productions. The high-COG female speaker has a mean COG of around 10,000 Hz for /s/, which is considerably higher than the mean COG of /s/ of the male speaker (∼7000 Hz) and the other female speaker (∼8900 Hz). She also has a narrower COG range (<2000 Hz) while the other two speakers have wider COG ranges (∼4000 Hz). These values are comparable with what has been observed in previous acoustic analysis studies for English fricatives, except that the high-COG female showed a higher COG distribution of /s/ than the mean value observed in the literature (e.g., Jongman et al., 2000).

Based on the acoustic data, the low-COG female resembles the male speaker to a greater extent than the high-COG female, in terms of the frequency distribution of spectral energy of sibilants.

3.3. Manipulation

All the spoken words were normalized to a consistent amplitude level and were manipulated into three kinds of stimuli:

Training stimuli. A set of 34 training stimuli containing ambiguous fricatives were synthesized in the male speaker's speech. Half of the 34 words contained /s/ word-medially and the other half contained /ʃ/ word-medially. A list of the 34 words can be found in Appendix B Table 3. They were adapted from Kraljic and Samuel (2006), with the /s/-words and /ʃ/-words matched in SUBTLEX lexical frequency count (Brysbaert & New, 2009) ($t = -0.02$, $p = 0.98$).

To create stimuli containing ambiguous fricatives, the 34 words were pronounced once with the correct sibilant and a second time with the incorrect sibilant (e.g., *initial* produced as [ɪ'nɪʃəl] first and [ɪ'nɪsəl] second). All of the correct and incorrect sibilants were then annotated in Praat TextGrid by hand. The proportion of the two sibilants in the same word frame (e.g., *in**i**t**ia**l* and *in**i**s**ia**l*) were cut out and blended at five steps of sibilant proportions, ranging from 0.3[s]0.7[ʃ] to 0.7[s]0.3[ʃ] with an increase of 0.1[s] and a decrease of 0.1[ʃ] at each interval. These blended fricatives were then spliced back into the lexical frame that originally contained the correct phoneme (to continue our example, the frame of *ini_ial* taken from the /ʃ/-containing pronunciation). A lexical decision task was then conducted to select the most ambiguous step of sibilant for each word frame to be used in the training phase. Participants

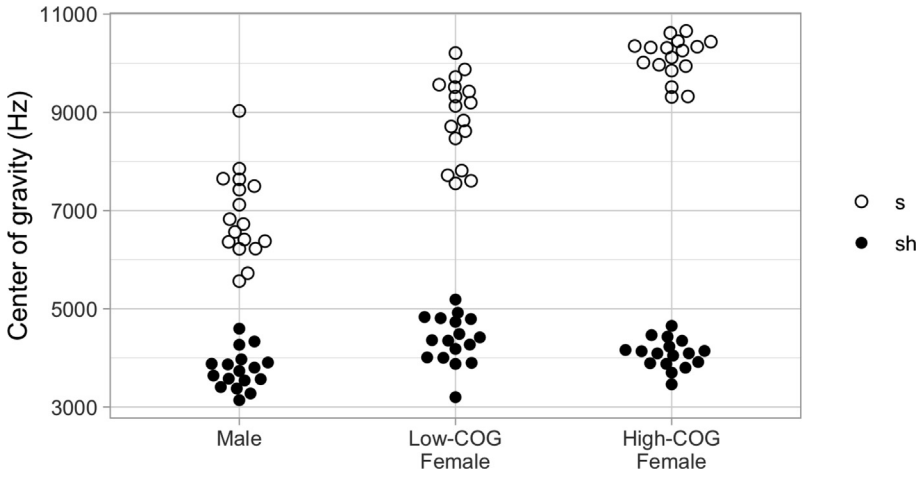

Fig. 1. The COG measures of sibilants from the three speakers in Hz.

were asked to judge whether what they heard was an English word or not, for each of the 34 lexical frames spliced with one of the five /s-ʃ/ steps (a total of 170 lexicality judgments). For each word, the mixture proportion that provided the most ambiguous (closest to 50%) categorization result was selected to be further used in the training materials for the perceptual learning experiment.

Test stimuli. A set of 35 test stimuli were synthesized in both the high-COG female's speech and the low-COG female speaker's speech. For each speaker, a continuum of five steps was generated by blending an original /s/ sound and a original /ʃ/ sound at five steps of sibilant proportions, ranging from 0.35 [s]0.65[ʃ] to 0.75[s]0.25[ʃ], with an increase of 0.1[s] and a decrease of 0.1[ʃ] at each interval. The five steps of sibilants were then spliced into seven word-frames of minimal pairs (*sake–shake, sign–shine, sell–shell, seat–sheet, same–shame, sigh–shy, self–shelf*), such that all the test words would form a different but real English word with the synthesized sound replaced by /s/ or /ʃ/.

Filler stimuli. We also recorded 33 filler words that did not contain sibilants anywhere in the word from each speaker. A list of filler words can be found in Appendix B Table 4.

3.4. Procedure

The experiment contains two kinds of blocks, training and test. Both types of blocks contained 51 spoken word identification trials.

A training block contained 34 training trials and 17 filler trials of words from the male speaker. The 34 training trials, composed of 17 trials of words containing phoneme /s/ and 17 containing /ʃ/, were designed to use lexical context to guide the perception of ambiguous sounds towards the intended direction. The crucial design that induces perceptual learning was that instances of one of the critical phonemes were natural whereas those of the other critical phoneme were acoustically ambiguous. Therefore, the /s/-favoring training condition contained ambiguous /s/ sounds in lexical /s/ contexts and natural /ʃ/ sounds in lexical /ʃ/ contexts, and the /ʃ/-favoring condition contained natural /s/ sounds in lexical /s/ contexts and ambiguous /ʃ/ sounds in lexical /ʃ/ contexts. For example, *initial* and *rehearsal* would be [ɪ'nɪʃəl] and [ɹɪ'hɚ?əl] in a /s/-favoring condition and [ɪ'nɪ?əl] and [ɹɪ'hɚsəl] in a /ʃ/-favoring condition. The provided response options are the correct word and one foil of a phonetically similar word. Importantly, they are not contrasted on the critical sound. In the above example, the options for /ɪ'nɪ?əl/ are *initiate* and *initial*, both supporting the perception of /ʃ/, and options for /ɹɪ'hɚ?əl/ are *rehearsal* and *reversal*, both pointing to /s/. The 17 filler trials did not contain any /s/ or /ʃ/.

A test block contained 35 test trials and 16 filler trials of words from one of the two female speakers. As described earlier, the 35 test trials included five steps on a /s/-/ʃ/ continuum for each of seven lexical frames of minimal pairs. Different categorizations of the critical sounds yielded two different words, which were provided as the two choices on that particular test trial. For example, for the stimulus /?eɪm/, listeners needed to choose between *same* and *shame* in response to what they heard.

Participants in the baseline conditions completed a test block with the speech of one of the two female speakers. Participants in the learning conditions first completed a training block with the male speaker's speech, which was manipulated to lexically favor either the perception of /s/ or /ʃ/ for ambiguous sibilants, before they completed a test block with one of the two female speakers' speech. The trial order was randomized within blocks for each participant, and the order of the two options was randomized for each trial.

3.5. Participants

A total of 179 participants were recruited to complete this experiment online. Of these, 48 were recruited through the online participant recruitment platform Prolific, and 131 were recruited from the University of Pennsylvania psychology subject pool. Participants were randomly assigned to the six experimental conditions. Table 1 shows the total number of participants in each condition, as well as the breakdown of numbers by gender, and the mean and standard deviation of the ages of the participants in each condition.

Table 1
Information of participants on the six conditions.

	Test with Low-COG Female		Test with High-COG Female	
	N (F, M)	Age mean (sd)	N (F, M)	Age mean (sd)
Baseline	31 (18, 13)	28.5 (9.0)	31 (25, 8)	19.7 (1.1)
Male /s/-favoring	34 (20, 14)	20.3 (2.5)	30 (20, 10)	19.5 (1.2)
Male /ʃ/-favoring	27 (17, 10)	24.5 (8.0)	26 (15, 11)	19.8 (1.0)

4. Results

In this section, we first present an acoustic analysis of the stimuli in different training conditions and by different test speakers after experimental manipulation. Then, according to the distributions of these stimuli, we lay out predictions about perceptual generalization according to the phonetic similarity account. Finally, we present the results of perceptual generalization along with the baseline categorization for each of the two female test speakers. Statistical analyses were conducted using the *R* Statistical environment version 4.1.0 (R Core Team, 2021); mixed-effects logistic regression was run using the *lme4* library version 1.1.27.1 (Bates, Mächler, Bolker, & Walker, 2014), and plots were created using *ggplot2* version 3.3.5 (Wickham, 2011).

4.1. Acoustic distributions of the training and test stimuli

In order to operationalize the phonetic similarity account in a testable way, we use the parameter of the spectral center of gravity as a stand-in for the spectral energy distributions of sibilants in our stimuli, which allows us to identify phonetic overlap between stimuli in the training and test conditions along the COG dimension. COG is converted from Hertz (Hz) to the psycho-acoustic mel scale to explore dynamics that are perceptually relevant. The mel scale is a logarithmic scale that transforms the linear Hz scale so that sounds of equal distance on the mel scale are perceived as equidistant. COG values in Hz were imported to the R programming environment and converted to the mel scale using the *mel* function from the *seewave* package (Sueur, Aubin, & Simonis, 2008).

In Fig. 2, each facet illustrates the COG values of fricatives from the 68 training stimuli of the male speaker (17 items $\times$ 2 phonemes $\times$ 2 conditions), which were identical across the two facets, and the COG values of five steps of test stimuli from either the high-COG female speaker or the low-COG female speaker. A solid outline of the half violin plot indicates that the sibilants in that condition were not manipulated, whereas a dashed outline of the half violin plot indicates that the sibilants was manipulated to be ambiguous between /s/ and /ʃ/, as has been detailed in Section 2.3.

The figure shows that training sibilants in the /ʃ/-favoring condition have overall higher COG values than those in the /s/-favoring condition. This is an expected outcome of the experimental manipulation, as we discussed in Section 2.3. As a result of these manipulations, the stimuli from the two training conditions also form different relationships with the two test continua in the acoustic space. For the low-COG female test speaker, the test continuum goes low enough that it overlaps in COG not only with the distribution of the higher training stimuli from the /ʃ/-favoring condition, but also with the distribution of the lower training stimuli from the /s/-favoring condition. By contrast, the test continuum for the high-COG female speaker overlaps only with the distribution of the higher stimuli in the /ʃ/-favoring condition, and not with the distribution of the /s/-favoring training condition.

According to our interpretation of the phonetic similarity account, we expect listeners to generalize their perceptual learning from the male training voice to the low-COG female test voice, regardless of which training condition they heard, because the test and training COG distributions overlap in both conditions. In contrast, because the high-COG female speaker's test continuum overlaps only with the /ʃ/-favoring training distribution, we expect listeners to only generalize from the male training stimuli in the /ʃ/-favoring condition. In the /s/-favoring condition, the distributions of the test and training voices are entirely non-overlapping and therefore provide no similarity basis for generalization.

4.2. Generalization to the low-COG female speaker

Fig. 3 shows the categorization results along the low-COG female speaker's /s–ʃ/ continuum by participants in the three experimental conditions. Participants in all groups exhibited more /s/-equivalent responses as a greater proportion of /s/ was blended in the test stimuli. Compared to the baseline condition, training with the male speaker's /s/-favoring stimuli boosted the average rate of /s/-equivalent responses, and training with the male speaker's /ʃ/-favoring stimuli reduced /s/-equivalent responses, on most sibilant steps of the continuum.

A logistic mixed-effects model was fitted to predict the categorization response for each trial (S = 1, SH = 0), with the fixed effects of Condition (treatment coded, reference: baseline condition) and Step (1–5, centered) in a two-way interaction, plus a random slope for Step by Subject (with correlated intercepts) and a random intercept by Word. The model shows a significant Step effect ($\beta = 0.15, p < 0.001$), meaning that higher proportions of /s/ blended in the test stimuli gave rise to more /s/-equivalent responses in the baseline condition. The Condition effect is significant for /ʃ/-favoring training ($\beta = -1.63, p < 0.001$) and marginally significant for /s/-favoring training ($\beta = 0.95, p = 0.06$), with coefficients consistent with the directions of the perceptual shifts. The interaction between Step and Condition is also significant for /ʃ/-favoring training ($\beta = -1.63, p < 0.001$) and marginally significant for /s/-favoring training ($\beta = 0.95, p = 0.06$), meaning that exposure to the male speaker's sibilants resulted in a shallower categorization boundary along the continuum than the one in the baseline condition. These results are not in favor of a speaker specificity account, which predicts no generalization across the training speaker and the test speaker. Instead, The above results are consistent with the prediction of the phonetic similarity account, because perceptual generalization has been

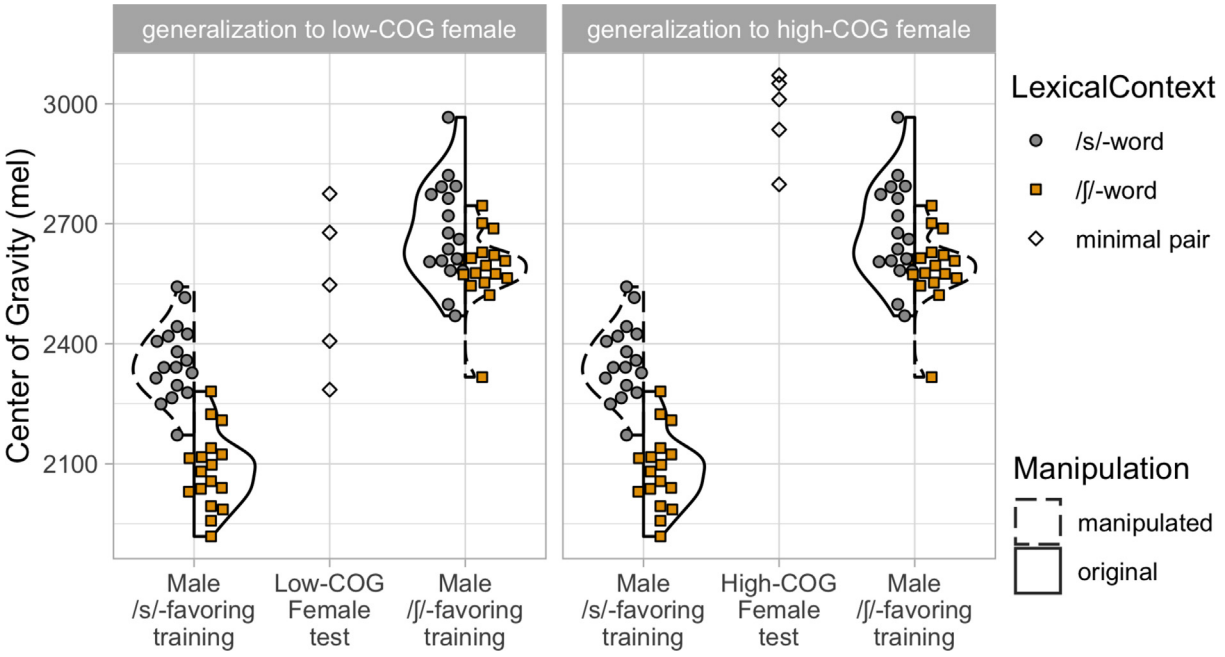

Fig. 2. Comparison of COG values of sibilants in the training stimuli from the male speaker with the test stimuli from the two female speakers (mel).

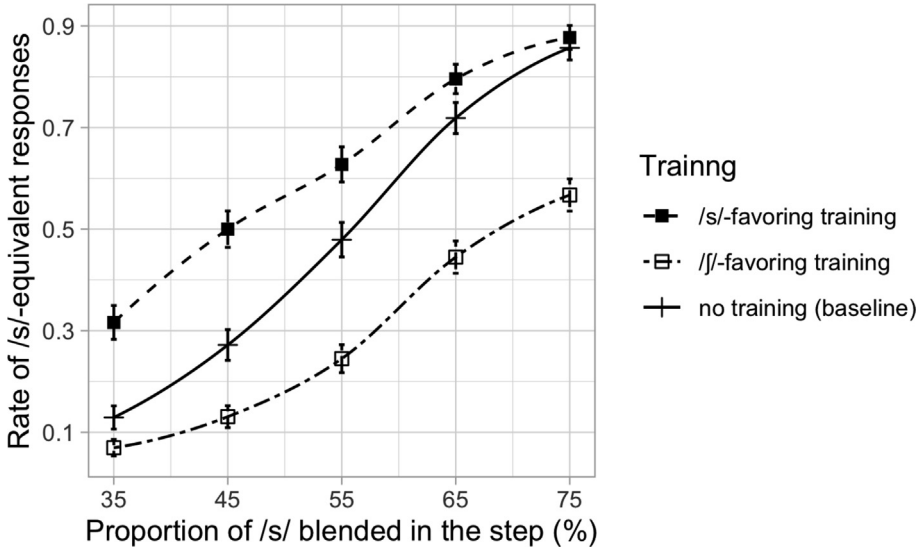

Fig. 3. Means and standard errors of /s/ response rates on the low-COG female speaker's /s/-/ʃ/ continuum in different experimental conditions.

observed across the male speaker and the low-COG female speaker, whose sibilant spectral energy distribution resembles the male speaker's, according to Section 3.2.

4.3. Generalization to the high-COG female speaker

Fig. 4 shows the categorization results along the high-COG female speaker's /s/-/ʃ/ continuum by participants with or without exposure to the male speaker's perceptually-biased speech. This time, we see that while training with the male speaker's /s/-favoring stimuli induced a certain amount of perceptual shift to the intended direction (boosted /s/ response rates), training with the male speaker's /ʃ/-favoring stimuli did not trigger a boundary shift in the perception of the high-COG female speaker's /s–ʃ/, given the overlapping between categorization results in the baseline condition and the male /ʃ/-favoring condition.

We again fit a logistic mixed-effects model to predict the categorization response for each trial (S = 1, SH = 0), with the fixed effects of Condition (treatment coded, reference: baseline) and Step (1–5, centered) in a two-way interaction, a random slope for Step by Subject (with the correlated intercept) and a random intercept by Word. The model showed a significant effect of Step ($\beta = 0.16, p < 0.001$), meaning that higher proportions of /s/ blended in the test stimuli led to more /s/-equivalent responses in the baseline condition. The Condition effect is marginally significant for /s/-favoring training ($\beta = -1.02, p = 0.05$) but not for /ʃ/-favoring training

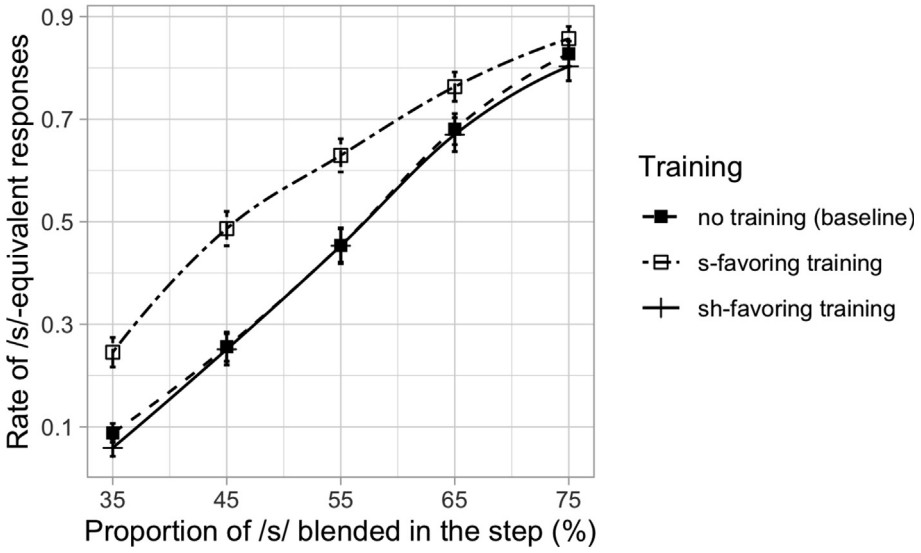

Fig. 4. Means and standard errors of /s/ response rates on the high-COG female speaker's /s/-/ʃ/ continuum in different experimental conditions.

($\beta = -0.29, p = 0.60$). In other words, training with the male speaker's /s/-favoring stimuli boosted the categorization of /s/ in the high-COG female's speech to a certain extent, but training with the male speaker's /ʃ/-favoring stimuli did not affect the categorization result, compared to the baseline condition. The interaction between Step and Condition is significant for /s/-favoring training ($\beta = -0.38, p = 0.03$) but not for /ʃ/-favoring ($\beta = -0.02, p = 0.90$).

The results of generalization to the high-COG female speaker are not consistent with the prediction of the phonetic similarity account, which predicted generalization in the /ʃ/-favoring condition and a lack of generalization in the /s/-favoring condition. Instead, we find exactly the opposite pattern: generalization in the /s/-favoring condition and lack of generalization in the /ʃ/-favoring condition.

5. Discussion

As a reminder, the question we set out to test was whether perceptual generalization is contingent on speaker identity or on the phonetic similarity between the critical training and test stimuli. Our experiment evaluated the relationship between the generalization of perceptual learning across multiple speakers and the phonetic characteristics of critical phonemes produced by those speakers. Specifically, we asked whether perceptual learning of sibilants from a male speaker would generalize to the perception of two female speakers' continua, one with sibilants with energy at lower frequencies as measured by the center of gravity (COG) and therefore bearing high similarity to the male speaker's production, and the other with sibilants of high COG and therefore bearing low similarity to the male speaker's production. We found that perceptual learning successfully generalized to the categorization of the low-COG female's sibilant continuum in both training directions. However, when listeners were tested on the /s/-/ʃ/ continuum of the high-COG female, whose sibilant productions are more distinct phonetically, generalization was only found in the /s/-favoring direction, but not the /ʃ/-favoring direction. As for our question of whether the generalization of perceptual learning would remain consistent for a given pair of speakers across experimental directions, then, the answer is "No." Our generalization results with the high-COG test speaker indicate that perceptual generalization does not necessarily remain consistent across speakers and can be contingent on the specific direction of the perceptual bias introduced in each experimental condition.

Since we do find generalization across different speakers in three of the four conditions, these results are clearly not consistent with a speaker specificity account. Our finding that perceptual learning generalizes successfully across the male speakers and the low-COG female speaker does seem to be consistent with what a phonetic similarity account would predict, namely, that perceptual learning generalizes across speakers whose critical phonemes overlap in the relevant phonetic dimension. However, the phonetic similarity account does not account for the one-sided perceptual generalization we found between the male speaker and the high-COG female speaker. The lack of generalization in the /ʃ/-favoring direction is not due to the absence of a learning effect in the /ʃ/-favoring condition in the first place, as evidenced by perceptual generalization in that condition to the low-COG continuum (i.e., with the other female test voice). It would also be difficult to attribute this lack of generalization to any kind of social dissimilarities between the male speaker and the high-COG female speaker, given that generalization *was* found between these two speakers in the /s/-favoring condition.

More specifically, our acoustic analysis of the critical items (Fig. 2) showed substantial overlap between the male training stimuli in the /ʃ/-favoring condition and the test stimuli from the high-COG female speaker in the dimension of spectrum center of gravity, which would allow for generalization to happen when evaluated in terms of phonetic similarity. By contrast, no phonetic overlap was found between the male training stimuli in the /s/-favoring condition and the test stimuli from the high-COG female speaker, which violates the requirement of phonetic similarity and therefore should have blocked generalization. However, our findings suggest exactly the opposite: we found that perceptual learning generalized marginally on the

/s/-favoring condition but not at all on the /ʃ/-favoring condition when tested on the high-COG female's continuum. This is an unexpected result that cannot be explained by either the speaker specificity or phonetic similarity accounts that we originally set out to compare.

Recent literature on perceptual learning has also noted asymmetries between opposite learning directions along the continuum of a phonemic contrast (Drouin, Theodore, & Myers, 2016; Zheng & Samuel, 2023). Drouin et al. (2016) reported more effective learning with /s/ than /ʃ/ by comparing ratings in a goodness judgement task. This asymmetry was also observed in Zheng and Samuel (2023), although their results indicated that the larger influence of /s/ only held for immediate testing and was gone after either an 24-h delay or a one-week delay. These studies are in line with our current results observed with the high COG speaker, showing that /s/-favoring training is more effective than /ʃ/-favoring training when there is an asymmetry observed. One possible explanation these authors put forward for why such an asymmetry might arise is that the phoneme that involves more phonetic variability will be more susceptible to learning (Drozdova et al., 2016; Drouin et al., 2016; Drouin & Theodore, 2018). This would explain why /s/-favoring training was more effective than /ʃ/ in the retuning of a /s/-/ʃ/ contrast. However, since this explanation is based on the phonetic properties of the training stimuli alone, it does not explain why the same set of training stimuli will trigger different results depending on the phonetics of the test stimuli. Therefore, an account based on the differential phonetic variability of different critical phonemes does not explain one of the main findings of our study; such an account would predict similarly asymmetrical generalization to the low-COG speaker, where we actually observed generalization in both directional conditions.

None of the existing accounts we are aware of will account for the pattern of generalization results that we have observed, leaving us without a ready-made explanation. Instead, we make a new proposal about a possible constraint on how perceptual learning generalizes. This proposal is rooted not just in the phonetic details of different voices, but in how those details get mapped to phonological contrasts. We refer to it as the **phonetics–phonology mismatch** account. In the following subsection, we lay out this account first with respect to the current experiment's stimuli, and then with respect to more general predictions it might make. We emphasize that the experiment we have reported here is not a *test* of the phonetics–phonology mismatch account; rather, we developed the account based on the outcome of this experiment. Future work could test its predictions more directly.

5.1. A phonetics–phonology mismatch account

The phonetics–phonology mismatch account that we propose requires us to distinguish between two kinds of information that listeners may learn in the training phase: a preference for one phoneme over the other, and a new boundary location in phonetic space. These two kinds of information may be thought of corresponding to "decision-making" and "category-representation" as outlined in Xie, Jaeger, and Kurumada (2023)'s computational framework for adaptive speech perception. These different mechanisms for behavioral change during speech perception are taken to arise from different levels of processing: the location of a new phoneme boundary involves remapping between perceptual and linguistic categories at the representation level, whereas the preference for one phoneme over the other induces an additional task-level bias that affects recognition. Either or both of these kinds of information could in principle be retained by the listener and put to use for categorization in the test phase, potentially giving rise to generalization. Our proposal is that both kinds of information are at play and must *match* to support successful generalization.

We start from the observation that one possible function of the manipulated training stimuli is to create a perceptual bias towards a particular phoneme category in the training speaker's phonological space. Essentially, one thing listeners may be learning in an /s/-favoring training condition is, "If you hear something you're not sure about in this sibilant space, you should tend to think it's an /s/" (and vice versa for /ʃ/). We will refer to /s/ in this example as the **identification-biased phoneme category**. Notice that our definition of the identification-biased phoneme is not tied to the particular phonetic values of the training speaker's stimuli, although of course it is learned from those stimuli; rather, it is a preference for one phoneme category over the other. In this light, the question of generalization is the question of whether listeners will transfer that phonological perceptual bias into their categorization behavior when they hear a new speaker.

However, another way to think about what happens in generalization is that the listener might attempt to transfer information they have just learned about the phonetic location of the boundary between /s/ and /ʃ/. We will refer to the categorization conclusion the listener would draw for some test item, if they imposed the training boundary location onto the test stimuli, as the **distribution-predicted phoneme category**. To be clear, we are not proposing that the listener in fact straightforwardly applies this boundary to the identification of the test stimuli (since that is obviously not the case), but rather that the evaluation of the test continuum against that boundary is one kind of directional pressure that the training stimuli can exert on the listeners' perception of the test stimuli.

Our proposal is that for perceptual learning to generalize across speakers, the distribution-predicted phoneme category for the ambiguous test stimuli needs to match the identification-biased phoneme category, for at least some of the test continuum. If the training stimuli are manipulated to favor the perception of one phoneme category, but the ambiguous test stimuli would exclusively be mapped onto a different phoneme category based on the phonetic distribution of the training stimuli, then we refer to it as a **phonetics–phonology mismatch** that interferes with generalization.

We now demonstrate how this proposal works more concretely, with reference to the results of the current study. With this account, we are able to explain why perceptual generalization does not occur to the high-COG female speaker's continuum in the /ʃ/-favoring condition in our experiment. Fig. 5 reproduces the COG measures of sibilant spectrum frequencies from Fig. 2, except that we added a blue dashed line for each training condition to indicate the location of the mean COG of the 34 training stimuli in that condition. We take the mean of the distribution of both phonemes to be a reasonable proxy for the boundary between the two phonemes. In other

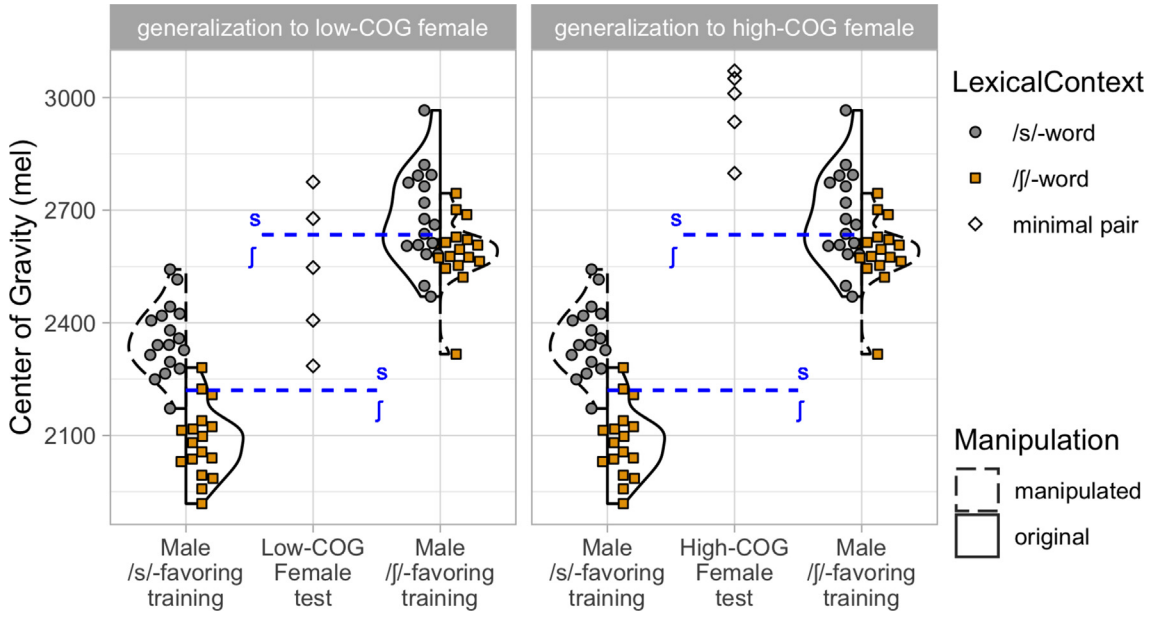

Fig. 5. COG values of sibilants in the training stimuli from the male speaker and the test stimuli from two female speakers. The dashed blue lines with IPA annotations indicate the mean COG of /s/ and /ʃ/ in different training conditions.

words, we assume that the COG values at the blue dashed lines can be interpreted as the locations of /s/-/ʃ/ boundary computed from the phonetic distribution of the training stimuli in different conditions: ambiguous sibilants with a COG value higher than the boundary are more likely to be an /s/, whereas ambiguous instances with a COG value lower than the boundary are more likely to be an /ʃ/.

In Fig. 5, the mean COG value of /s/ and /ʃ/ in the /s/-favoring condition is 2220 mel whereas the mean COG value of /s/ and /ʃ/ in the /ʃ/-favoring condition is 2635 mel. These values set up different phonetic standards for the categorization of /s/-/ʃ/ between these conditions: In the /ʃ/-favoring condition, sibilants should be categorized as /s/ only when they have a COG value higher than 2635 mel, whereas in the /s/-favoring condition, sibilants with a lower COG value can also be categorized as /s/, because the threshold has been lowered to 2220 mel. The different boundary locations imposed by these two training conditions would trigger different categorizations for fricative tokens with COG values that fall within the range of 2220–2635 mel.

Now we consider the application of these boundary locations to the ambiguous instances from the test speakers. For the low-COG female speaker (left facet), in the /s/-favoring training condition, all the test stimuli would fall into the category of /s/ according to the phonetic boundary imposed under that condition, because they all have a COG value larger than 2220 mel. The distribution-derived phoneme category for the test stimuli and the identification-biased phoneme category in the training condition are both /s/, and therefore the perceptual shift towards /s/ can be generalized. In the /ʃ/-favoring condition, with the COG threshold of /s/ increased to 2635 mel, 3 out of the 5 ambiguous instances (steps 1–3, with COG values of 2285 mel, 2407 mel, 2547 mel) would be categorized as /ʃ/ according to the boundary location at 2635 mel. At least for these three instances, the distribution-predicted phoneme category is the same as the identification-biased phoneme category (both /ʃ/), which allows perceptual learning to generalize in this condition.

For the high-COG female speaker, we see a similar configuration in the /s/-favoring training condition. All the test stimuli would fall into /s/ because their COG values are all higher than 2220 mel, which is the location of the phoneme boundary in that condition. The distribution-predicted phoneme category for the test stimuli is consistent with the identification-biased phoneme category in that experimental condition (both /s/), allowing for the generalization of perceptual learning across speakers. By contrast, in the /ʃ/-favoring training condition, all the ambiguous instances still fall into the category of /s/, because their COG measures are all higher than the boundary location computed from the /ʃ/-favoring stimuli (2635 mel). However, this distribution-derived phoneme category for the test stimuli (/s/) is at odds with the perceptually favored phoneme category by the training stimuli (/ʃ/). We suggest that, faced with this clash in how the two different kinds of learned information would extend to the new voice, the listener stops attempting to generalize what they have learned to the new voice. Thus, the phonetics–phonology mismatch blocks the generalization of perceptual learning.

5.2. Generalizing and extending the phonetics–phonology mismatch account

Importantly, if the account we have just sketched out is on the right track, it should allow us to derive not only specific predictions for idiosyncratic combinations of training and test voices, but also broader generalizations about types of voice combinations and training directions that are more likely to give rise to generalization. Here we return to two points from Section 2.3: the intrinsic phonetic consequences of different manipulation directions, and the covariation of gender and sibilant frequency.

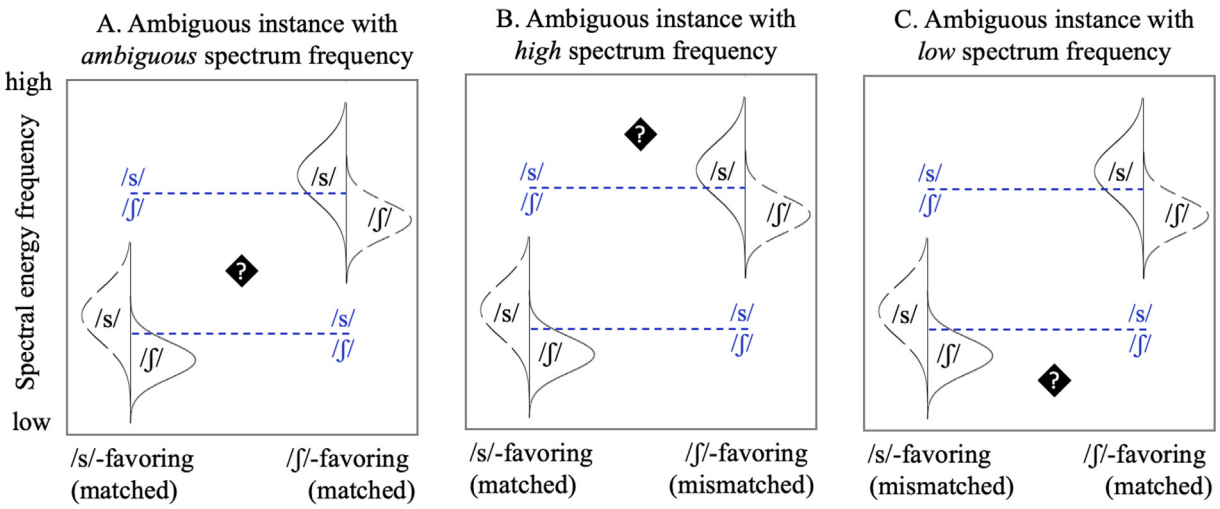

Fig. 6. A schema of phonetics–phonology mismatch in the generalization of /s/-/ʃ/ perceptual learning. Solid line distributions represent naturally produced instances and dashed line distributions represent manipulated ambiguous instances lexically coerced to different phoneme interpretations. Question marks stand for an ambiguous instance in the test phase.

Fig. 6 shows how the phonetic locations of /s/ and /ʃ/ in different perceptual learning conditions are intrinsic to stimulus manipulation methods in a typical perceptual learning paradigm. An /s/-favoring condition coerces ambiguous instances to an /s/ identification while keeping natural instances of /ʃ/, so that listeners learn to categorize instances in the middle of the continuum into /ʃ/. Similarly, an /ʃ/-favoring condition coerces ambiguous instances to /ʃ/ while keeping natural instances of /s/. As a result, an /s/-favoring training condition would end up with sibilants of spectral energy at lower frequencies while an /ʃ/-favoring training condition would have sibilants of higher spectral frequency. When the speaker remains the same in the training phase and the test phase, it is a given that the ambiguous instance falls into the phonetic region that can be categorized in different ways as intended in different conditions. This situation is illustrated in Facet A, where the identity of the ambiguous instance is sensitive to the shift of phonetic distributions of the training stimuli. At this point, there is no phonetics–phonology mismatch between the training stimuli and the test instance.

Now consider the case where the training speaker and the test speaker vary in their mapping between sibilant categories and the frequency distribution of spectral energy, perhaps because they are speakers of different genders. The combination of a male training speaker and a female test speaker can give rise to a situation as in Facet B (mirroring the right facet of Fig. 5), where sibilants of the training speaker have greater spectral energy at lower frequencies than those of the test speaker. Accordingly, the ambiguous instance from the test speaker would locate at the high-frequency end of the training speaker's phonetic space. In the /ʃ/-favoring condition in Facet B, although the training stimuli are manipulated to have energy at higher spectral frequencies to allow for ambiguous instances from the training speaker to be categorized as /ʃ/, the spectral frequency distributions are still not high enough so that the ambiguous instance from a female test speaker could become mapped to /ʃ/. In Facet B, we have a mismatch between the identification bias in the training speaker's phonetics–phonology mapping and the distribution-predicted phoneme category of the test speaker's ambiguous instance in the /ʃ/-favoring condition. Note that this problem does not exist in the /s/-favoring condition, where the test ambiguous instance is mapped onto the phoneme that the experimental manipulation is intended to favor.

Similarly, the combination of a female training speaker and a male test speaker might give rise to the situation in Facet C, where the training speakers have sibilants with energy at higher spectral frequencies than the test speaker. As a result, the spectral energy frequency of the ambiguous stimuli of the test speaker is distributed at the lower end of the training speaker's phonetic space. This could potentially bring about a phonetics–phonology mismatch in the /s/-favoring perceptual learning condition. In the /s/-favoring condition, the spectral frequency distributions of the sibilants are lowered to make more room for categorizing the training speaker's ambiguous instance as /s/. However, given the test speaker has an intrinsically lower (spectral) COG than the training speaker, the phonetic location of the test speaker's ambiguous instance still corresponds to /ʃ/ by the phonetics–phonology mapping of the training stimuli in the /s/-favoring condition, creating a mismatch between the distribution-predicted phoneme category and the identification-biased phoneme category in that condition. This mismatch does not exist in the /ʃ/-favoring condition in Facet C, where the test speaker's high-frequency ambiguous instance is also phonetically mapped onto the training speaker's /ʃ/.

This proposal has broader implications, then, for when perceptual learning of this contrast would typically be expected to generalize given the covariation between sibilants and speaker gender. Given the information that female speakers tend to produce sibilants with energy at higher frequencies than male speakers, the phonetics–phonology mismatch account makes the direction-specific prediction that when trained with a male speaker and tested with a female speaker, generalization would be more likely to occur in the /ʃ/-favoring perceptual learning condition but be minimal in the /s/-favoring condition; when trained with a male and tested with a female, generalization would be more likely to occur in the /s/-favoring perceptual learning condition but be minimal in the /ʃ/-favoring condition. Fig. 6 illustrates how these predictions are derived.

Of course, this is only a broad generalization, since the phonetic properties of different speakers' voices are idiosyncratic and gender is far from the sole factor shaping sibilant production. Our own demonstration of different outcomes with different female test voices should make this abundantly clear. However, as a population-level generalization, it has the potential to be of interest in understanding connections between the individual linguistic behavior of perceptual learning and the community linguistic phenomenon of sound change (Tamminga et al., 2020).

Lastly, the phonetics–phonology mismatch account proposed in this paper could also be useful for making predictions about the perceptual generalization of other linguistic categories in addition to sibilants, as long as robust covariation between speech production and speaker identity can be identified. Similar cases include vowel formants and F_0 contrasts, for which the acoustic realizations are also conditioned on speaker gender. For example, given that male speakers produce lower vowel formants than female speakers for the same vowel category, the phonetics–phonology mismatch constraint would predict that perceptual learning that favors the perception of a higher vowel (such as /ɪ/) would not (or only minimally) generalize from a female speaker to a male speaker, whereas the perceptual learning that favors the perception of a lower vowel (such as /ɛ/) would not generalize from a male speaker to a female speaker. Similarly, in the perceptual learning of lexical tones, we predict that perceptual learning that favors the perception of a high tone would not generalize from a female speaker to a male speaker, whereas perceptual learning that favors the perception of a low tone would not generalize from a male speaker to a female speaker. These predictions remain to be tested by future studies.

Although the phonetics–phonology mismatch constraint and the phonetic similarity account make different predictions about generalization in the current experiment, the formulation of these two accounts is ultimately driven by the same logic and intuition: The more similar the production of the critical phonemes are between two speakers, the more likely it is that they have similar phonetics–phonology mappings, and the less likely it is that a phonetics–phonology mismatch as we have described would occur. Variability in speech production results in inconsistency in the mapping between the phonetic distributions and phonological categories, which might block perceptual generalization from happening. In this respect, the phonetics–phonology mismatch account can be understood more as a refinement of the phonetic-similarity account than as a competitor.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Acoustic Measures

See Table 2.

Appendix B. Stimulus Lists

See Tables 3 and 4.

Table 2
Acoustic measures of critical phonemes from the three speakers.

speaker	Male		Low-COG Female		High-COG Female	
phoneme	/s/	/ʃ/	/s/	/ʃ/	/s/	/ʃ/
spectral variance	2377	1958	2479	1991	1901	1593
skewness	0.155	1.99	-0.858	1.38	-0.220	2.15
kurtosis	1.71	7.23	3.58	5.39	5.36	8.66
duration (ms)	109	108	120	122	134	155
normalizedDur	15.5%	15.2%	15.9%	17.0%	16.1%	18.3%

Table 3
Word list of trials and foils in the training phase.

	/ʃ/ word	freq	foil	/s/ word	freq	foil
1	compensate	124	condensate	ambition	273	inhibition
2	dinosaur	203	dining-set	beneficial	40	artificial
3	embassy	397	embarrassing	brochure	97	butcher
4	episode	627	webisode	commercial	829	financial
5	eraser	51	harasser	crucial	234	cruel
6	falsetto	15	falsehood	efficient	253	effective
7	faucet	73	flawless	evaluation	225	valuation
8	hallucinate	15	deracinate	glacier	38	Grayshott
9	legacy	256	legally	graduation	500	substitution
10	medicine	1744	medical	impatient	206	impacted
11	obscene	176	obscuring	initial	325	essential
12	parasite	126	parasol	negotiate	342	negation
13	peninsula	70	Pennsylvania	official	1224	optimal
14	personally	1870	personality	parachute	162	paragon
15	pregnancy	334	presidency	publisher	230	publicly
16	reconcile	58	gracile	refreshing	187	infringing
17	rehearsal	635	reversal	vacation	1673	vocation
		Mean freq.: 419.8			Mean freq.: 402.2	

Table 4
Word list of fillers in the training and test phases.

airline	amongst	anvil	average	banana	beloved	buffalo
dragonfly	earning	eyebrow	gable	gargoyle	honey	iguana
January	jewelry	journal	lonely	marina	enamel	Nepal
nothing	racoon	raven	ribbon	row	runaway	thumbnail
verify	village	volleyball	vugar	waffle		

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

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

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavioral Research Methods, 41*(4), 977–990. Nov.

Chodroff, E., & Wilson, C. (2020). Acoustic–phonetic and auditory mechanisms of adaptation in the perception of sibilant fricatives. *Attention, Perception, & Psychophysics, 82*(4), 2027–2048.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658.

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *The Journal of the Acoustical Society of America, 144*(2), 1089–1099.

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America, 140*(4), EL307–EL313.

Drozdova, P., van Hout, R., & Scharenborg, O. (2016). Processing and adaptation to ambiguous sounds during the course of perceptual learning.

Durvasula, K., & Nelson, S. (2018). Lexical retuning targets features. In Proceedings of the Annual Meetings on Phonology. Vol. 5.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & psychophysics, 67*(2), 224–238.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America, 84*(1), 115–123.

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

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.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178.

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15.

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition, 121*(3), 459–465.

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

Lev-Ari, S., & Peperkamp, S. (2014). Do people converge to the linguistic patterns of non-reliable speakers? perceptual learning from non-native speakers. In *10th International Seminar on Speech Production: Satellite Workshop on "Interpersonal coordination and phonetic convergence* (pp. 261–264).

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America, 125*(6), 3962–3973.

Nirgianaki, E. (2014). Acoustic characteristics of greek fricatives. *The Journal of the Acoustical Society of America, 135*(5), 2964–2976.

Nissen, S. L., & Fox, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *The Journal of the Acoustical Society of America, 118*(4), 2570–2578.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 539.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics, 45*, 91–105.

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

Shadle, C. H. & Mair, S .J. (1996). Quantifying spectral characteristics of fricatives. In Proceeding of fourth international conference on spoken language processing. ICSLP'96. Vol. 3. IEEE (pp. 1521–1524).

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In KONVENS (pp. 14–26).

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics, 18*(2), 213–226.

Tamminga, M., Wilder, R., Lai, W., & Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology, 5*, 90–122.

Tjaden, K., & Turner, G. S. (1997). Spectral properties of fricatives in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 40*(6), 1358–1372.

Weatherholtz, K. (2015). *Perceptual learning of systemic cross-category vowel variation* Ph.D. thesis. The Ohio State University.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics, 3*(2), 180–185.

Wikse Barrow, C., Włodarczak, M., Thörn, L., & Heldner, M. (2022). Static and dynamic spectral characteristics of swedish voiceless fricatives. *The Journal of the Acoustical Society of America, 152*(5), 2588–2600.

Włodarczak, M. (2022). Swedish fricatives. (Last viewed August 26, 2023). doi: 10.5281/zenodo.7248299.

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.*.

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: General, 150*(11), e22.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America, 143*(4), 2013–2031.

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 200.

Zheng, Y., & Samuel, A. G. (2023). Flexibility and stability of speech sounds: The time course of lexically-driven recalibration. *Journal of Phonetics, 97*, 101222.