

1 The effects of exposure variability and similarity on
2 talker-independent adaptation to Spanish-accented
3 English

4 Holly A. Zaharchuk^{a,b,*}, Janet G. van Hell^a

^a*Department of Psychology and Center for Language Science The Pennsylvania State University*

^b*Department of Speech Language and Hearing Sciences and Institute for Brain and Cognitive Sciences University of Connecticut*

5 **Abstract**

6 The influence of a talker's first language (L1) on the production of their
7 second (L2) can complicate the task of word recognition. Nevertheless, pre-
8 vious research has shown that listeners demonstrate remarkable speed and
9 plasticity in adapting to L2-accented talkers. Moreover, listeners are able to
10 generalize this learning to novel L2-accented talkers. There are two compet-
11 ing explanations for these talker-independent adaptation effects: one argues
12 that increasing variation during exposure facilitates generalization to a novel
13 test talker (exposure-to-variability hypothesis), while the other argues that
14 increasing acoustic-phonetic similarity between the exposure and test talkers
15 facilitates generalization (similarity-based hypothesis). We conducted three
16 experiments to compare the effects of exposure variability and exposure-test
17 similarity on generalization. During exposure, monolingual English partici-
18 pants performed an auditory lexical decision task with multisyllabic English
19 tokens spoken by three Spanish-accented talkers. After exposure, we tested
20 performance on a novel Spanish-accented talker with a cross-modal match-
21 ing task (Experiments 1 and 2) or a primed cross-modal lexical decision task
22 (Experiment 3). Exposure variability increased both identity and competi-
23 tor priming, consistent with a relaxation of phonetic categorization crite-
24 ria. Exposure-test similarity decreased competitor priming without chang-
25 ing identity priming, while exposure uniformity increased identity priming
26 without changing competitor priming, providing some evidence of phonetic
27 recalibration in these conditions. Our results provide equivocal support for
28 the similarity-based hypothesis and against the exposure-to-variability hy-

*Corresponding author

Email address: holly.zaharchuk@uconn.edu (Holly A. Zaharchuk)

29 pothesis.

30 *Keywords:* Adaptation, Generalization, Perceptual learning,
31 Foreign-accented speech, Word recognition

1 Data availability

2 The data and analyses included in this manuscript can be found on the
3 Open Science Framework (<https://osf.io/z3q6v>). The full list of experimen-
4 tal stimuli and the post-experiment questionnaire can also be found in the
5 repository.

6 1. Introduction

7 The production of a single speech sound varies both diachronically within
8 an individual talker and synchronically across groups of talkers. Listeners
9 must accommodate both types of variation in the speech signal in order to
10 achieve successful comprehension. Perceptual adaptation is the process of
11 detecting the *patterns* of variation in the speech signal and applying them
12 during comprehension. This learning process allows listeners to map unfa-
13 miliar, variable, or secondary acoustic cues onto phonetic category represen-
14 tations. Previous research has shown that perceptual adaptation to specific
15 cue-category mappings can transfer from one talker to another (e.g., Kraljic
16 & Samuel, 2006, 2007; Reinisch & Holt, 2014). In addition, the benefits of
17 exposure to an L2 accent can transfer to a new talker with the same accent.
18 However, there is disagreement in the literature about the conditions under
19 which adaptation generalizes across talkers. Two different factors have been
20 argued to facilitate generalization: (1) *variability* during exposure (Baese-
21 Berk et al., 2013; Bradlow & Bent, 2008) and (2) *similarity* between the
22 exposure and test talkers (Xie et al., 2021; Xie & Myers, 2017). This paper
23 reports three experiments that directly compared variability and similarity
24 during exposure to Spanish-accented English. The overarching goal is to
25 understand how listeners use acoustic-phonetic information to accommodate
26 linguistic variation.

27 1.1. L2-accented speech

28 During L2 acquisition, a speaker’s L1 phonetic inventory shapes the pro-
29 duction of L2 speech sounds (Flege & Bohn, 2021; Nagle & Baese-Berk,
30 2022). We focused on three contrasting stop consonant pairs found in both
31 English and Spanish: /p/-/b/ (*park* vs. *bark*), /t/-/d/ (*tune* vs. *dune*), and
32 /k/-/g/ (*coal* vs. *goal*). Within each pair, one member differs from the other
33 in terms of voicing (voiceless vs. voiced). Voicing refers to the status of
34 the glottis, which comprises two folds of tissue located in the larynx at the

top of the neck. Constricting the vocal folds causes them to vibrate, resulting in voiced speech sounds; by contrast, opening the vocal folds limits vibration, resulting in voiceless speech sounds. The primary cue to word-initial stop consonant voicing in both English and Spanish is voice onset time (VOT), which is the interval between the release of the closure of the articulators, called the burst, and the beginning of voicing (Cho & Ladefoged, 1999; Chodroff et al., 2019; Lisker & Abramson, 1964). Listeners are highly sensitive to the probabilistic distributions of VOT across voiced and voiceless stops in their language (Clayards et al., 2008).

The associations between VOT and stop consonants vary cross-linguistically (Chodroff et al., 2019; Lisker & Abramson, 1970). In Spanish, lead (negative) VOTs correspond to voiced stops and short lag VOTs correspond to voiceless stops (Vicente, 1986; Williams, 1977). For example, the /b/ in Spanish *barco* (boat) would be approximately -98 ms on average, while the /p/ in Spanish *parque* (park) would be approximately 16 ms on average (values calculated from Chodroff et al., 2019). By contrast, in English, short lag VOTs correspond to voiced stops and long lag VOTs correspond to voiceless stops (Chodroff & Wilson, 2017). For example, the /b/ in English *bark* would be approximately 9 ms on average, while the /p/ in English *park* would be approximately 60 ms on average (values calculated from Chodroff et al., 2019). As a result, the VOT for Spanish /p/ is a better fit for the contrasting English /b/ phoneme category than for the corresponding English /p/ phoneme category. The overlap in VOT between Spanish /p/ and English /b/ is illustrated in Figure 1, with probability densities constructed from the data presented in Chodroff et al. (2019).

To illustrate how cross-language transfer in production introduces ambiguity in perception, consider an L1 Spanish speaker who is learning English as an L2. She already has an established /p/-/b/ contrast in her L1 Spanish that is associated with VOT. The /p/-/b/ contrast in her L2 English is associated with the same cue in the same direction, such that VOTs for /p/ are longer than those for /b/. As a result, the English /p/-/b/ contrast will be relatively easy to integrate into her existing phonological system compared to, for example, certain vowel contrasts in English (Baigorri et al., 2019). The same will be true for both the /t/-/d/ and /k/-/g/ contrasts. However, assimilating the L2 English /p/ to the L1 Spanish /p/ means that the English phoneme will be produced like the Spanish phoneme. That is, L2 English /p/ will be produced with a short lag VOT (similar to Spanish /p/). This will create a perceptual problem from an L1 English listener's perspective, since

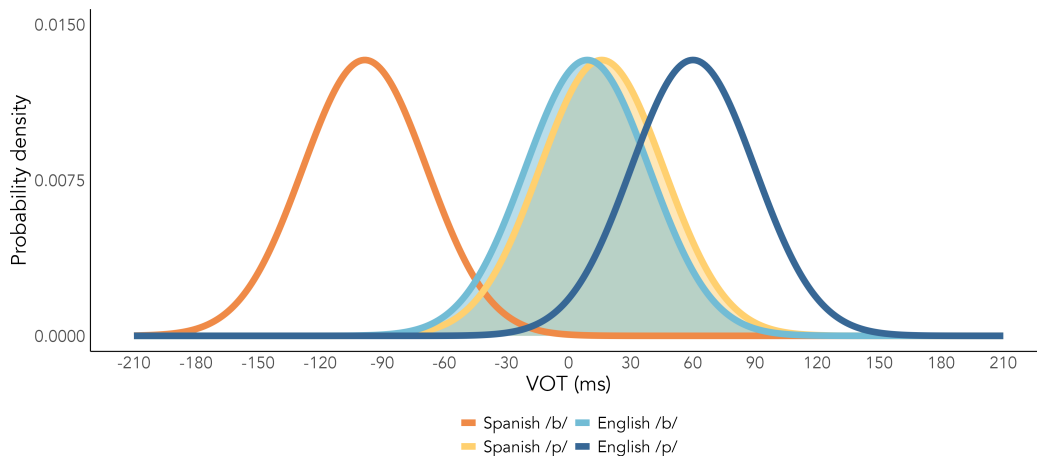


Figure 1: Overlapping VOT distributions for Spanish voiceless stops and English voiced stops.

73 short lag VOTs are associated with English /b/, not /p/; without support-
 74 ing contextual information, listeners may confuse Spanish-accented English
 75 /p/ with /b/ (e.g., *park* perceived as *bark*). To summarize, cross-language
 76 transfer in Spanish-accented English can create ambiguity between voiced
 77 and voiceless stops (Flege & Eefting, 1987). Listeners are adept at resolving
 78 such ambiguities through perceptual adaptation.

79 1.2. Adaptation to L2-accented speech

80 Listeners rapidly fine-tune their phonetic categories in response to L2-
 81 accented speech (Bent & Baese-Berk, 2021). A typical experimental design
 82 features an exposure phase, in which participants gain experience with a
 83 speech pattern, followed by a test phase. Talker-specific adaptation is the re-
 84 sult of hearing the same talker during exposure and test (Clarke & Garrett,
 85 2004; Xie et al., 2017). Talker-independent adaptation, or generalization,
 86 is the result of exposure to a *type* of talker rather than exposure to a *spe-*
 87 *cific* talker (Alexander & Nygaard, 2019; Bradlow et al., 2023; Bradlow &
 88 Bent, 2008; Sidaras et al., 2009; Tzeng et al., 2016; Xie et al., 2018, 2021).
 89 Overall, experience with L2-accented talkers improves perception of novel
 90 L2-accented talkers (Baese-Berk et al., 2013; Bieber & Gordon-Salant, 2022;
 91 Reinisch & Holt, 2014; Tzeng et al., 2024; Wittman et al., 2013). The pat-
 92 tern of results for talker-independent adaptation becomes more complicated
 93 when considering different types of exposure to an accent. Studies often com-

94 pare single-talker exposure, in which one talker produces all of the stimuli
95 for a condition, to multi-talker exposure, in which several different talkers
96 produce the stimuli for a condition. The literature is divided as to whether
97 multi-talker exposure is necessary for generalization (Baese-Berk et al., 2013;
98 Bradlow et al., 2023; Bradlow & Bent, 2008; Xie et al., 2021; Xie & Myers,
99 2017).

100 Two experiments highlight the necessity of multi-talker exposure for talker-
101 independent adaptation. In Bradlow and Bent (2008, Experiment 2), par-
102 ticipants completed a sentence transcription task during both exposure and
103 test, and performance was measured in terms of sentence recognition ac-
104 curacy. There were four exposure conditions: talker-specific, multi-talker,
105 single-talker, and control. Participants in the talker-specific condition were
106 trained directly on the Mandarin-accented English test talker, while those
107 in the multi-talker condition were trained on five novel Mandarin-accented
108 English talkers. Participants in the single-talker condition were exposed to
109 just one of the Mandarin-accented English talkers from the multi-talker con-
110 dition. Control training featured five L1-accented English talkers. Training
111 was followed by two post-tests: one featured a novel talker with a familiar
112 L2 accent (Mandarin) and the other featured a novel talker with an unfamil-
113 iar L2 accent (Slovakian). Performance on the Mandarin-accented English
114 test talker was higher in the talker-specific and multi-talker conditions than
115 in the single-talker and control conditions, which did not differ from one
116 another. By contrast, performance on the Slovakian-accented English test
117 talker did not differ between conditions. These results suggest that exposure
118 to multiple talkers with the same L1 highlights the features that characterize
119 a specific L2 accent.

120 Baese-Berk et al. (2013) provided evidence that exposure to multiple
121 L2-accented talkers with *different* L1s also facilitates generalization. They
122 used the same stimulus materials and procedures as Experiment 2 in Bradlow
123 and Bent (2008) in order to compare a multi-accent exposure group directly
124 to the multi-talker exposure and control groups from the previous study.
125 During multi-accent training, participants were exposed to five L2-accented
126 English talkers, each of whom had a different L1: Thai, Korean, Hindi,
127 Romanian, and Mandarin. Multi-accent and multi-talker exposure yielded
128 higher performance on the Mandarin-accented test talker than control expo-
129 sure. Multi-accent exposure also generalized to the Slovakian-accented test
130 talker, while multi-talker and control exposure did not. Taken together with
131 the results of Bradlow and Bent (2008), these findings suggest that exposure

132 to multiple L2-accented talkers facilitates talker-independent adaptation to
133 a greater extent than exposure to a single L2-accented talker.

134 By contrast, another set of studies found that single-talker exposure is
135 sufficient for generalization (Bradlow et al., 2023; Xie et al., 2021). Xie et al.
136 (2021) specifically sought to replicate Bradlow and Bent (2008) with one key
137 difference: counterbalancing the assignment of talkers to the exposure and
138 test phases across participants. This change allowed them to address a po-
139 tential confound between the type of exposure (multi- vs. single-talker) and
140 the specific talkers used for exposure and test. They replicated the original
141 finding that multi-talker exposure to Mandarin-accented English facilitates
142 generalization more than control exposure. However, contrary to Bradlow
143 and Bent (2008), they found that single-talker exposure also facilitated gen-
144 eralization more than control exposure. When comparing multi- and single-
145 talker exposure, the difference in performance depended on the particular
146 combination of exposure and test talkers. Bradlow et al. (2023) also observed
147 asymmetries in performance between single- and multi-talker exposure when
148 rotating four talkers, each with different L1s, through the exposure and test
149 roles. Interestingly, they observed these asymmetries even when the test
150 talker was an exposure talker. Overall, Bradlow et al. (2023) concluded that
151 single-talker exposure can be sufficient for generalization when the talkers
152 have the right combination of exposure-test similarity and baseline intelligi-
153 bility.

154 Xie and Myers (2017) found that generalization is strongly influenced by
155 talker-specific features. Since the design of the present study was based on
156 that of Xie and Myers (2017), we will describe the paradigm in detail here.
157 During exposure, participants performed an auditory lexical decision task
158 with multisyllabic items produced by five Mandarin-accented English talkers
159 (Experiment 1). In two follow-up studies, participants performed the same
160 exposure task with only one Mandarin-accented English talker (Experiments
161 2 and 3). Critically, participants were randomly assigned to one of two ex-
162 posure groups that featured different sets of real words on critical trials.
163 Experimental exposure included /d/-final words (e.g., *overload*), while con-
164 trol exposure included words without /d/ in any position (e.g., *animal*). The
165 exposure manipulation focused on word-final /d/ because it is perceptually
166 confusable with /t/ in Mandarin-accented English (e.g., *seed* perceived as
167 *seat*). During test, participants performed a primed cross-modal lexical deci-
168 sion task, in which auditory primes produced by a novel Mandarin-accented
169 talker were followed by visual targets. On critical trials, the visual targets

170 were either /d/-final real words or their /t/-final minimal pairs (e.g., seed or
171 seat). The priming effect, or the difference in response times between trials
172 with matching /d/-final auditory primes (e.g., seed) and those with unre-
173 lated auditory primes (e.g., fair), indexed the amount of talker-independent
174 adaptation.

175 The multi-talker exposure results indicated that listeners shifted their
176 phonetic boundaries with exposure to word-final /d/ in line with other work
177 on generalization. Specifically, the experimental group exhibited larger prim-
178 ing effects for matching /d/-final targets than the control group. This implies
179 that auditory *seed* became a better prime for visual *seed* after exposure to
180 disambiguating lexical contexts, suggesting that exposure increased activa-
181 tion of the intended word. The experimental group also exhibited larger
182 priming effects for matching /d/-final targets than for competing /t/-final
183 targets, while the magnitude of these priming effects did not differ in the
184 control group. This implies that auditory *seed* became a worse prime for vi-
185 sual *seat* after exposure to disambiguating lexical contexts, suggesting that
186 exposure decreased lexical activation of the unintended word. In other words,
187 exposure reduced competition between /d/- and /t/-final minimal pairs. If
188 listeners in the experimental group had simply expanded their phonetic cat-
189 egory for /d/, the priming effects for /d/-final targets would still have been
190 larger than those in the control group (indicating increased activation of the
191 matching targets), but the priming effects for /t/-final targets would have
192 also been larger (indicating increased activation of the competing minimal
193 pairs). Overall, the results showed that exposure to word-final /d/ in disam-
194 biguating lexical contexts like *overload* reshaped listeners' phonetic categories
195 for /d/ and /t/, such that the auditory prime *seed* was not only perceived
196 more like the matching word *seed*, but also perceived less like its minimal
197 pair *seat*.

198 The single-talker exposure results demonstrated that multi-talker expo-
199 sure is not necessary for generalization. Two of the five exposure talkers from
200 the main experiment were selected for single-talker exposure: one who was
201 the most similar to the test talker on three key acoustic measures of word-
202 final /d/ and one who was the least similar. When the exposure talker was
203 similar to the test talker, the experimental group exhibited larger priming
204 effects for /d/-final targets than for /t/-final targets, while the control group
205 did not. By contrast, when the exposure talker was dissimilar to the test
206 talker, there were no differences between exposure groups or target types.
207 These findings suggest that exposure to a single talker with similar word-

208 final /d/ realizations to the test talker reduced competition between /d/-
209 and /t/-final minimal pairs. Moreover, this reduction in lexical competition
210 with a single similar exposure talker was comparable to the one observed in
211 the main experiment with multiple exposure talkers. Together, the results of
212 Xie and Myers (2017) show that the correspondence between exposure and
213 test talkers at the acoustic-phonetic level is critical for understanding the
214 observed patterns of generalization across studies.

215 1.3. *Competing hypotheses for generalization*

216 There are two competing explanations for why multi-talker exposure may
217 or may not provide additional benefits for generalization over single-talker ex-
218 posure. The **exposure-to-variability hypothesis** posits that L2-accented
219 talkers exhibit similarities in production that differ from L1-accented norms
220 (Baese-Berk et al., 2013). The exposure-to-variability hypothesis also posits
221 that among L2-accented talkers with the *same L1*, cross-language influence
222 shifts the relations between acoustic cues and phonetic categories similarly
223 across talkers. Among L2-accented talkers with *different L1s*, typological
224 features that are unique to the L2 lead to similarly accented realizations
225 regardless of the L1. Multi-talker exposure thus allows listeners to abstract
226 away from the peculiarities of any given talker and home in on these common-
227 alities. For example, a listener encountering one unfamiliar Spanish-accented
228 English talker may not know whether their short lag VOTs are specific to
229 that talker or characteristic of the Spanish accent. By contrast, a listener
230 encountering multiple unfamiliar Spanish-accented English talkers at once
231 would not only hear that short lag VOTs are common across the talkers, but
232 also that other accent features (e.g., vowel height) exhibit covariation with
233 voicing (Clayards, 2017). In Bradlow and Bent (2008), multi-talker expo-
234 sure outperformed single-talker exposure, suggesting that training on multi-
235 ple L2-accented talkers (with the same L1) allowed listeners to separate the
236 characteristic features of an accent from the idiosyncratic features of a talker.
237 The results of Bradlow and Bent (2008) support the exposure-to-variability
238 hypothesis, but do not align straightforwardly with the effects observed by
239 Xie and Myers (2017), Xie et al. (2021), or Bradlow et al. (2023).

240 The **similarity-based hypothesis** posits that acoustic-phonetic overlap
241 between the exposure and test talkers, rather than variability during expo-
242 sure, facilitates generalization (Xie et al., 2021). In Xie and Myers (2017),
243 listeners exhibited comparable talker-independent adaptation effects after
244 both single- and multi-talker exposure to Mandarin-accented realizations of

word-final /d/, which is perceptually confusable with /t/ (e.g., *seed* vs. *seat*). Critically, both exposure conditions contained a Mandarin-accented talker with similar word-final /d/ acoustics to the test talker. Xie et al. (2021) also demonstrated equivalent generalization effects from single- and multi-talker exposure to Mandarin-accented English. They argue that exposure to multiple talkers merely increases the likelihood that listeners will encounter a cue distribution that is relevant for adapting to the test talker. For example, multi-talker training on Spanish-accented speech is more likely to include at least one talker with short lag VOTs than single-talker training. Bradlow and Bent (2008)’s multi-talker exposure condition may have included talkers who were more similar to the test talker than their single-talker exposure condition, which would confound talker similarity and exposure to variability. Xie et al. (2021) and Bradlow et al. (2023) addressed this confound by counterbalancing the combinations of exposure and test talkers across participants. Thus, the similarity-based hypothesis may account for the differential effects of single- and multi-talker exposure in perceptual adaptation studies. The present study was designed to distinguish between the roles of variability and similarity in talker-independent adaptation. To summarize, the similarity-based hypothesis focuses on specific cue-category mappings, while the exposure-to-variability hypothesis focuses on the covariation between cues.

1.4. Comparing variability and similarity

As discussed above, the results of Bradlow and Bent (2008) and Baese-Berk et al. (2013) support the exposure-to-variability hypothesis, while those of Xie and Myers (2017) and Xie et al. (2021) support the similarity-based hypothesis. The results of Bradlow et al. (2023) fell in between, with adaptation depending greatly on the specific combinations of talkers. We argue that the ideal adapter framework provides a link between these two hypotheses and can account for these conflicting findings (Kleinschmidt, 2019; Kleinschmidt & Jaeger, 2015). Under this theory, listeners are sensitive to the probability distributions of cue-category mappings—like the ones shown in Figure 1 for the acoustic cue VOT over the phonetic categories /p/ and /b/—and either create new or updated existing mental representations of these distributions through statistical learning. The specificity of these cue-category representations, called generative models, depends on two factors: how much information is gained about a cue distribution and how well a category predicts a cue by using a particular grouping. An ideal adapter will

282 structure her generative models according to the most “informative” and
283 “useful” socio-indexical groupings (Kleinschmidt, 2019). For L2-accented
284 speech, these groupings may be structured according to each talker (talker-
285 specific) or to the L2 accent shared by the talkers (talker-independent). The
286 socio-indexical structure of cue-category mappings is critical for reconciling
287 the exposure-to-variability and similarity-based hypotheses.

288 The exposure-to-variability hypothesis argues that increasing variability
289 during exposure increases systematic covariation among relevant cues, en-
290 abling listeners to separate the idiosyncratic features of a talker from the
291 common features of an accent. Similar arguments have been made regarding
292 high variability phonetic training paradigms for non-native contrasts (Brekel-
293 mans et al., 2022). To restate this hypothesis in terms of the ideal adapter
294 framework, covariation among individual exposure talkers promotes the de-
295 velopment of a robust talker-independent model (or refinement of an ex-
296 isting one). In turn, this talker-independent model guides adaptation to
297 the novel test talker. This exposure-to-variability perspective assumes that
298 talker-independent models are better for generalization than talker-specific
299 models.

300 By contrast, the similarity-based hypothesis argues that increasing the
301 similarity between the cue-category mappings of the exposure and test talkers
302 facilitates generalization. Translating this hypothesis into the ideal adapter
303 framework, listeners develop robust talker-specific models during exposure.
304 During test, listeners use the model that provides the most information about
305 the test talker’s cue distributions and most readily predicts their individual
306 cue values to guide adaptation. This similarity-based perspective assumes
307 that talker-specific models are best for generalization because they provide
308 precise information about cue-category mappings.

309 Overall, the specificity of a listener’s set of generative models may explain
310 the effects of variability and similarity on generalization. The present study
311 tests these two hypotheses in order to probe the mechanisms underlying
312 adaptation to L2-accented speech. The overarching goal is to understand
313 how the L1 speech recognition system learns the patterns of cross-language
314 influence in L2 speech production.

315 1.5. *Present study*

316 How do listeners generalize their experience with L2-accented speech? On
317 the one hand, the *structure* of exposure may be the primary driver of talker-
318 independent adaptation. Variability has been the primary focus of structural

inquiries (Baese-Berk et al., 2013), with number of talkers being the key manipulation (Bieber & Gordon-Salant, 2022; Bradlow & Bent, 2008; Choi & Perrachione, 2019; Xie et al., 2021). On the other hand, the *content* of exposure may be the key to effective generalization. At a high level, training on the relevant L2 accent tends to facilitate talker-independent adaptation (Alexander & Nygaard, 2019; Bradlow et al., 2023; Clarke & Garrett, 2004; Xie et al., 2018). At a detailed level, exposure to the relevant acoustic-phonetic features of an L2 accent explains generalization beyond shared L1 (Reinisch & Holt, 2014; Sidaras et al., 2009; Xie & Myers, 2017). Together, a shared L1 and common acoustic-phonetic features create similarity between exposure and test talkers that facilitates adaptation. The present study builds on this body of work by investigating generalization of L2-accented speech. The key difference between our study and earlier work is the operationalization of variability and similarity. Specifically, we increased the precision with which variability was implemented (cf. Baese-Berk et al., 2013) and expanded the scope of similarity (Bradlow et al., 2023; cf. Xie & Myers, 2017) in order to better delineate the exposure-to-variability and similarity-based hypotheses. The goal was to clarify some of the inconsistent findings in the literature for talker-independent adaptation to L2-accented speech (Bent & Baese-Berk, 2021).

Regarding variability, previous studies have almost exclusively investigated this factor as a comparison between single-talker and multi-talker exposure (cf. Baese-Berk et al., 2013; Bradlow et al., 2023). However, multi-talker exposure reduces the amount of experience with any given talker relative to single-talker exposure. From the larger perceptual adaptation literature, we know that listeners are highly sensitive to acoustic-phonetic mappings and use them to adapt to new talkers (e.g., Kraljic & Samuel, 2006). We also know that there is a high degree of within- and between-talker variability in both L1- and L2-accented speech production (Wade et al., 2007; Xie & Jaeger, 2020). Depending on the idiosyncrasies of a given exposure talker, altering the amount of experience with this talker may benefit or impede generalization. These asymmetries in talker-specific exposure may explain some of the inconsistent findings in the literature (Bradlow et al., 2023; Xie et al., 2021; Xie & Myers, 2017). Moreover, exposure to multiple talkers increases the sources of covariation to account for, which in turn increases the difficulty of accounting for them during exposure. If we assume, as the exposure-to-variability hypothesis does, that experience with covariation is critical for generalization, this lack of clarity obscures how and why variability

might facilitate generalization. In the present study, we used a single group of talkers with the same L1 to control the type and amount of experience with each talker across levels of variability.

Regarding similarity, previous studies have taken two general approaches. One focuses on the L1 of the exposure and test talkers, with similar talkers sharing an L1 and dissimilar talkers having different L1s (e.g., Bradlow & Bent, 2008). Another focuses on the acoustic-phonetic features of the exposure and test talkers, with similar talkers having greater overlap in their distributions than dissimilar talkers (e.g., Xie & Myers, 2017). To operationalize similarity, both approaches use different talkers for each exposure condition (rendering the same problem we discussed above for multi-talker vs. single-talker exposure). As a result, we cannot know whether listeners are “learning a talker or learning an accent” or learning both (Xie & Myers, 2017). In other words, this design confounds low-level talker-specific effects and high-level accent-specific effects. Here, we manipulated the stimuli listeners encounter during exposure rather than the talkers in order to maintain consistency across levels of similarity. Moreover, we directly crossed the factors of similarity and variability in the same design, which, to our knowledge, has not yet been done. Overall, these manipulations provide a novel way to address the ongoing debate between the exposure-to-variability and similarity-based hypotheses for generalization. A summary of the experimental design in relation to our research questions is presented in the next section. The data and analyses for all three experiments can be found on the Open Science Framework (<https://osf.io/z3q6v>).

2. General methods

We conducted three experiments with an exposure-test design to understand how different kinds of experience with Spanish-accented speech change listeners’ VOT-stop mappings. The exposure phase established the comparison between the similarity-based and exposure-to-variability hypotheses of talker-independent adaptation. Participants performed an auditory lexical decision task with a between-subjects manipulation that exposed them to different sets of multisyllabic real words and pseudowords (Xie & Myers, 2017). In Experiment 1, there were six between-subjects conditions. In Experiments 2 and 3, two of these conditions were removed, and changes to the pseudowords were made to improve learning.

392 The test phase assessed the effects of each type of exposure on perception.
 393 In Experiments 1 and 2, participants performed a cross-modal matching task
 394 with a within-subjects manipulation that tested a common set of monosyl-
 395 labic real words. In Experiment 3, participants performed a cross-modal
 396 primed lexical decision task that better assessed learning from exposure.
 397 Both tasks featured auditory primes and visual targets and were adapted
 398 from Xie and Myers (2017). Across experiments and phases, participants
 399 heard four Spanish-accented talkers—three during exposure and one during
 400 test—with the talker’s role counterbalanced across participants. We also
 401 conducted a norming study to select the items for both tasks, which can be
 402 found in Appendix A.

403 2.1. Stimulus design

404 To clearly describe the overall experimental design, it is necessary to first
 405 describe the design of the stimuli. Stimuli were grouped by onset phoneme
 406 (e.g., *park* has the onset phoneme /p/). There were three **experimental**
 407 **onset** groups: critical, competitor, and control. Critical onsets were voiceless
 408 stops (/p/, /t/, and /k/), competitor onsets were voiced stops (/b/, /d/,
 409 and /g/), and control onsets were voiceless fricatives (/f/, /s/, and /θ/).
 410 Across the experimental onset groups, phonemes were also grouped (roughly)
 411 according to their place of articulation: labial (/p/, /b/, and /f/), alveolar
 412 (/t/, /d/, and /s/), and postalveolar/velar (/k/, /g/, and /θ/). These will
 413 be referred to as the **cross-experimental onset** groups. There was one
 414 **filler onset** group: /m/, /n/, /l/, /r/, /h/, and /w/. Filler onsets were
 415 also grouped by their place of articulation: nasal (/m/ or /n/), alveolar (/l/
 416 or /r/), and other back consonants (/h/ or /w/). These will be referred to
 417 as the **within-filler onset** groups. Finally, each cross-experimental onset
 418 group was paired with one of the within-filler onset groups according to
 419 frontness/backness: front (/p/, /b/, /f/, /m/, and /n/), mid (/t/, /d/, /s/,
 420 /l/, and /r/), and back (/k/, /g/, /θ/, /h/, and /w/). These will be referred
 421 to as the **cross-condition onset** groups.

422 2.2. Exposure phase design

423 **Exposure similarity** was operationalized as the relation between the
 424 experimental onsets encountered during exposure and the critical onsets en-
 425 countered during test. In Experiment 1, there were three levels of Simi-
 426 larity: Direct, Indirect, and Control. Experiments 2 and 3 included the
 427 Direct and Indirect levels only. Each level refers to the type of information

428 participants received about Spanish-accented voiceless stops. In Direct con-
429 ditions, participants were exposed directly to critical onsets (e.g., *peanut*).
430 In Indirect conditions, participants were exposed to competitor onsets (e.g.,
431 *beehive*), thereby gaining experience with the shifted VOT continuum that
432 they would encounter during test. In Control conditions, participants were
433 exposed to control onsets (e.g., *football*), thereby gaining general experience
434 with Spanish-accented English but not with stop VOTs. This design allowed
435 the talkers to remain the same across levels of Similarity. The top of Figure
436 2 provides examples of the real words that comprised each level of Similarity;
437 the examples were selected in order to illustrate the three cross-experimental
438 onset groups that were included across levels (front, mid, and back/other).

439 **Exposure variability** was operationalized as the relation between onset
440 phonemes and exposure talkers. There were two levels of Variability: Invari-
441 ant and Variant. Each level refers to the type of experience with each talker.
442 In Invariant conditions, listeners heard each of the three exposure talkers
443 produce one onset out of the three in an experimental group (e.g., Direct-
444 Invariant: Talker A produced *peanut*, *palace*, and *pianist*; Talker B produced
445 *terminal*, *tiara*, and *textile*; and Talker C produced *kingdom*, *counter*, and
446 *kayak*). This means that all of the words with a given onset were produced
447 by only one talker. In Variant conditions, listeners heard each of the three
448 exposure talkers produce all three onsets in an experimental group (e.g.,
449 Direct-Variant: Talker A produced *peanut*, *terminal*, and *kingdom*; Talker B
450 produced *palace*, *tiara*, and *counter*; and Talker C produced *pianist*, *textile*,
451 and *kayak*). This means that one third of the words with a given onset were
452 produced by each talker. This design allowed the items and talkers to re-
453 main the same between levels of Variability. The two levels of Variability are
454 depicted at the top of Figure 2, with green arrows/boxes illustrating Vari-
455 ant talker assignments and purple arrows/boxes illustrating Invariant talker
456 assignments.

457 2.3. Test phase design

458 There were three levels of Target: Identity, Competitor, and Unrelated.
459 Each level refers to the type of visual target that followed each auditory
460 prime. Identity targets exactly matched the auditory primes (e.g., *park-*
461 *park*). Competitor targets were the minimal pairs of the auditory primes
462 (e.g., *park-bark*). Unrelated targets only shared vowels with the auditory
463 primes (e.g., *park-wand*). Differential performance on these three conditions
464 indexed learning from exposure. The test phase is depicted at the bottom of

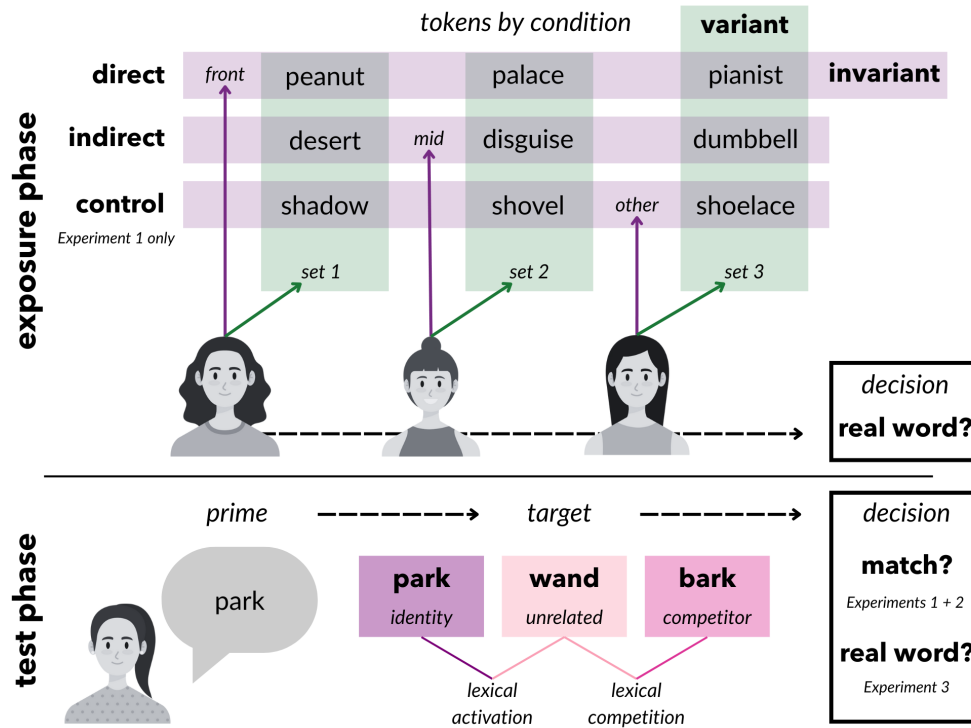


Figure 2: Cross-experiment design.

Figure 2 with the auditory prime *park* and its corresponding visual targets. In Experiments 1 and 2, participants made an explicit decision about whether the auditory prime matched the visual target. In Experiment 3, participants made a decision about whether the visual target was a real word or not, with the preceding auditory stimulus acting as a prime for the lexical decision.

2.4. Predictions

The exposure-to-variability and similarity-based hypotheses make different predictions about the effects of each type of exposure on test. The test task in Experiments 1 and 2 measures explicit judgments about the match/mismatch between auditory primes and visual targets, while the test task in Experiment 3 measures more implicit priming effects. In both cases, performance on Identity and Competitor targets relative to Unrelated targets indexes changes in perception as a result of exposure. Smaller differences in accuracy and RTs between Competitor and Unrelated targets

479 (*park-bark* vs. *park-wand*) would be evidence of decreased lexical competi-
480 tion. Larger differences in accuracy and RTs between Identity and Unrelated
481 targets (*park-park* vs. *park-wand*) would be evidence of increased lexical ac-
482 tivation. Exposure conditions that facilitate perceptual adaptation should
483 elicit both of these patterns, which together indicate that listeners have fine-
484 tuned their phonetic categories to accommodate Spanish-accented stops. The
485 presence of only one of these patterns would provide some evidence of per-
486 ceptual adaptation. A reversal of the matching/priming effects (i.e., larger
487 differences in performance between Competitor and Unrelated targets and
488 smaller differences between Identity and Unrelated targets) would indicate
489 a general relaxation of phonetic category boundaries rather than perceptual
490 recalibration (Zheng & Samuel, 2020).

491 The similarity-based hypothesis predicts a main effect of Similarity, with
492 evidence for perceptual adaptation increasing from Control to Indirect to
493 Direct exposure. This prediction follows directly from the ideal adapter
494 framework, such that both talker-specific and talker-independent generative
495 models of VOT-stop distributions can generalize to a test talker; what mat-
496 ters is how similar the exposure model is to the test distribution. As the
497 name implies, Direct exposure provides the most direct experience the VOT
498 distributions that characterize voiceless stops in Spanish; however, Indirect
499 exposure also provides experience with these distributions that listeners may
500 use to generalize to the novel talker. If listeners are sufficiently sensitive to
501 the covariation between voiced and voiceless stop VOTs within talkers, the
502 similarity-based hypothesis would also predict comparable patterns of gen-
503 eralization from both Direct and Indirect exposure (e.g., Kraljic & Samuel,
504 2006).

505 The exposure-to-variability hypothesis predicts a main effect of Variabil-
506 ity, with more robust perceptual adaptation after Variant exposure than after
507 Invariant exposure. Under this hypothesis, exposure to covariation between
508 cues and categories encourages the formation of talker-independent genera-
509 tive models. This implies that such models have more utility for categorizing
510 critical onsets than any talker-specific models (Kleinschmidt, 2019). The
511 exposure-to-variability hypothesis does not predict effects of Similarity on
512 test performance.

513 2.5. Talkers

514 Four female Spanish-accented English talkers from Mexico recorded the
515 experimental stimuli (Talkers 1-4). In addition, a female L1-accented En-

516 glish talker from the US recorded the practice stimuli (Talker 0). All four
517 L2-accented talkers reported that they grew up speaking Spanish with their
518 caregivers and began acquiring English in traditional classroom settings. Ad-
519 ditional language background information and recording details can be found
520 in Appendix B.

521 We will briefly describe the VOT distributions of each talker here and
522 refer the reader to Appendix B for details. Talker 4 produced the most
523 Spanish-like VOTs across places of articulation, with shorter means than
524 the other talkers. Talker 3 produced overall longer VOTs than the other
525 talkers, which even fell outside L1 US English norms. Talkers 1 and 2 were
526 the most similar to one another, with Talker 1 aligning the most with L1
527 English norms. Overall, each talker exhibited a unique pattern of VOT-stop
528 distributions.

529 **3. Experiment 1: Investigating differences in generalization from** 530 **exposure to Spanish-accented stops versus fricatives**

531 *3.1. Methods*

532 *3.1.1. Participants*

533 We recruited 296 participants through the online platform Prolific. Par-
534 ticipants provided implied consent in line with Penn State IRB policies and
535 were compensated \$6 after completing the experiment, which took approxi-
536 mately 30 minutes. The experiment was available to individuals whose Pro-
537 lific user profiles aligned with the following eligibility criteria: between 18 and
538 40 years of age, located in the US at the time of the study, English as their
539 first and only fluent language, normal hearing, normal or corrected-to-normal
540 vision, and without a history of language-related disorders. The Prolific user
541 profiles for each participant also included sex (two options, select one) and
542 race/ethnicity information (four options, select all that apply).

543 The eligibility criteria were cross-checked with the responses to a post-
544 experiment questionnaire (see Section 3.1.4), and any ineligible participants
545 were removed from further analysis ($N = 9$). We also removed participants
546 who reported that their L1 or L2 was Spanish, with self-rated proficiency
547 greater than or equal to 3/5 in any languages other than English, with self-
548 rated proficiency less than 5/5 in English, or whose place of origin was not
549 the US ($N = 15$). This left 287 eligible participants.

550 Finally, we removed participants with poor data quality from further
551 analysis ($N = 15$). Poor data quality was defined as: accuracy statistically

indistinguishable from or significantly below chance on experimental real words in the exposure task, zero correct experimental trials in any of the three conditions in the test task, or fewer than 50% of experimental trials with correct responses and reaction times between 50 and 2500 ms in either task (see Section 3.1.4 for details). This left 257 participants for analysis (Age: $M = 31$, $SD = 6$, Min = 18, Max = 40; Sex: Female = 121, Male = 135, Prefer not to say = 1; Race: Asian = 5, Black = 43, Multiple selected = 21, Other = 5, White = 182, Not provided = 1). Our *a priori* power analysis for the 2 x 3 factorial design (Variability x Similarity) indicated that we needed 27 participants per condition, which we achieved ($1-\beta = 0.8$, $\alpha = .05$, and Cohen's $f = 0.25$).

We recruited 46 additional participants through Prolific to complete the experiment without the exposure phase. Removing ineligible participants left 38 participants. Removing participants with low data quality left 37 participants for analysis (Age: $M = 31$, $SD = 5$, Min = 21, Max = 40; Sex: Female = 19, Male = 18; Race: Asian = 2, Multiple selected = 5, Other = 2, White = 28). All aspects of recruitment were the same for this group as for the main group of participants.

3.1.2. Stimuli

In total, there were 936 auditory items and 648 visual items across the exposure and test tasks. The final set of exposure items included 360 real words (24 per onset) and 360 pseudowords (24 per onset). The final set of auditory primes included 216 real words (24 per onset). The final set of visual targets included 648 real words (3 per auditory prime). An additional set of six real words and six pseudowords with filler onsets (one per onset per word type) was selected for practice. Six additional primes with filler onsets (one per onset) were selected for practice. Two additional unrelated targets were also selected for practice. Stimulus selection is described in Appendix C.

3.1.3. Experimental lists

Combining Similarity and Variability created six between-subjects conditions: Direct-Variant, Direct-Invariant, Indirect-Variant, Indirect-Invariant, Control-Variant, and Control-Invariant. In addition, one group of participants did not receive any exposure prior to test; this will be referred to as the Test-only condition. The 288 filler items were the same across conditions and evenly divided by word type—real word and pseudoword—as well as by onset—/m/, /n/, /l/, / /, /h/, and /w/. The 144 experimental items were

588 evenly divided by onset and word type, with onset differing by level of Sim-
589 ilarity: Direct (critical: /p/, /t/, and /k/), Indirect (competitor: /b/, d/,
590 and /g/), and Control (control: /f/, /s/, and / /). As a result, there were
591 24 items per onset per word type.

592 The assignment of talkers to items differed by level of Variability. In
593 Invariant conditions, talkers were assigned by cross-condition onset group:
594 front (/p/, /b/, /f/, /m/, and /n/), mid (/t/, /d/, /s/, /l/, and / /), and
595 back/other (/k/, /g/, / /, /h/, and /w/). In other words, one talker was
596 assigned to front onsets, one to mid, and one to back/other. Within a given
597 level of Similarity, this means that each talker produced items with one out
598 of the three experimental onsets and two out of the six filler onsets. For
599 Variant conditions, one third of each of the items of a given onset and word
600 type (8) was randomly assigned to one of three sets: set 1, set 2, or set 3.
601 Each set was assigned to one talker.

602 To counterbalance which talkers were assigned to which items across par-
603 ticipants, each Variant item set was paired with one Invariant onset group
604 to create three assignment groups: group 1 (front with set 1), group 2 (mid
605 with set 2), and group 3 (back/other with set 3). In addition, we also coun-
606 terbalanced the talkers between the exposure and test phases. To do this, we
607 added group 4 (test) and rotated the four talkers across these four assignment
608 groups in a Latin square design. Overall, this resulted in 24 experimental
609 lists for the exposure phase, one for each combination of exposure condition
610 (6) and talker assignment (4).

611 Regardless of exposure condition, all participants heard the same 216 au-
612 ditory primes during test. One third (72) were critical items divided evenly
613 by onset—/p/, /t/, and /k/—and two thirds (144) were filler items divided
614 evenly by onset—/m/, /n/, /l/, / /, /h/, and /w/. Each onset was divided
615 evenly by target type—Identity, Competitor, and Unrelated. This resulted
616 in eight items per onset per target type. Three experimental lists were cre-
617 ated to counterbalance the combinations of auditory prime and visual target
618 across participants. Since the test talker was also counterbalanced across
619 participants, this resulted in 12 experimental lists for the test phase, one for
620 each combination of prime-target pair (3) and talker assignment (4).

621 Exposure practice was presented in Talker 0's voice for all participants.
622 Test practice was presented according to the participant's talker assignment
623 and Variability condition. Participants in the Test-only condition completed
624 the test practice in Talker 0's voice.

3.1.4. Tasks

The experiment included a headphone check, exposure task, test task, and post-experiment questionnaire. All aspects of the experiment were conducted online using Pavlovia. The headphone check, exposure task, and test task were created in PsychoPy Builder (Peirce et al., 2019). The post-experiment questionnaire was built using Pavlovia’s survey platform.

The headphone check followed the anti-phase tone test procedure from Woods et al. (2017). On each trial, participants listened to three pure tones, one of which was out of phase with the other two. Participants indicated which of the three tones was the quietest by selecting the appropriate button on the screen: tone 1, tone 2, or tone 3. There were two trials per response for a total of six trials, which were presented in random order. Participants using well-functioning headphones should have easily perceived the anti-phase tone as the quietest, while those using loudspeakers should not. Participants completed the task at most two times.

The exposure phase featured the auditory lexical decision task from Xie and Myers (2017). On each trial, participants indicated whether an auditory stimulus was a real English word or not by pressing the *d* or *k* key on their keyboard. One of the two response-key relations—real-*d* or real-*k*—was assigned randomly to each participant. Participants completed 12 practice trials followed by 432 main trials presented in random order. Half of the practice trials (6) and half of the main trials (216) required real word responses.

The test phase featured a cross-modal matching task adapted from the primed cross-modal lexical decision task in Xie and Myers (2017). On each trial, participants first heard a real word (auditory prime) and then saw a real word written on the screen (visual target). They indicated whether the visual target matched the auditory prime or not by pressing the *d* or *k* key on their keyboard. The assignment of responses to keys was carried over from the exposure phase, with *match* responses mapped to the same key as *real word* responses in the exposure task. Participants completed six practice trials followed by 216 main trials presented in random order. Half of the practice trials (3) and one third of the main trials (72) required match responses.

The post-experiment questionnaire included two sets of items. The first set of items related to the talker from the test task. The second set of items included questions about the participant’s own language background and demographics. This second set of items was used to confirm the participant’s

eligibility as described in Section 3.1.1. Further details about the questionnaire items and descriptive statistics can be found in Appendix D.

3.1.5. Procedure

Eligible participants accessed the experiment through Prolific. Once a participant began the experiment, they were randomly assigned to one of the experimental lists. First, they performed the headphone check. If the participant achieved fewer than five correct trials out of six, they completed the task again. Participants who failed the headphone check a second time were not allowed to continue; instead, they were redirected to Prolific and asked to return their submission. After passing the headphone check, participants continued to the exposure task and completed the practice (participants in the Test-only condition continued straight to the test task). If a participant scored 50% or lower on the exposure practice, they were not allowed to continue; instead, they were redirected to Prolific and asked to return their submission. After successfully completing the practice session, the participant performed the exposure task. Next, the participant continued to the test task, where they completed the practice session and the main task. Immediately following the test task, participants continued to the post-experiment questionnaire. Once the questionnaire was complete, the participant was redirected to Prolific and received compensation.

3.1.6. Analysis approach

Data processing and analysis were conducted with R version 4.5.1 (R Core Team, 2022). Filler items were removed before proceeding with the following data processing pipeline and therefore were not included in any of the statistical analyses.

Prior to analyzing the exposure task data, we removed responses with RTs less than 50 ms ($N = 11$; 0.03%). RT was calculated from the onset of the word. For the exposure task, analyses were conducted on the accuracy data only.

Prior to analyzing the test task data, we removed responses with RTs less than 50 ms ($N = 4$; 0.02%). RT was calculated from the presentation of the visual target. For the test task, analyses were conducted on both the accuracy and RT data. For the RT analyses, we filtered for correct responses. We used the *robustbase* package to calculate adjusted boxplot statistics (appropriate for such skewed distributions as reaction time) to detect responses with reaction times outside the upper and lower fences (Hubert & Vandervieren,

2008). We detected and removed outliers by Target type (total $N = 898$; 4.24%). The upper and lower fence calculations included data from Experiment 2 in order to avoid using different outlier criteria for the Test-only group between experiments.

Mixed-effects models were fitted to trial-level data with the *lme4* package (Bates et al., 2015). Generalized linear mixed-effects models with a binomial distribution and logit linking function were used to analyze binary accuracy (1,0). Generalized linear mixed-effects models with an Inverse Gaussian distribution and identity linking function were used to analyze RT (see Lo & Andrews, 2015). Models were fit with the full random effects structure that was relevant to the task (see below). To test the significance of a fixed effect, Type-III analysis-of-deviance tables were calculated and Wald chi-square tests conducted with the *car* package (Fox & Weisberg, 2019).

Exposure task analyses modeled the fixed effects of Variability, Similarity, Word type, and their interactions on accuracy. The two levels of Variability and Word type were sum contrast-coded. The three levels of Similarity were simple-coded with Direct as the reference level. Scaled VOT, trial, and word frequency were included as covariates. Participant, talker, and item were included as random intercepts. Word type was included as an uncorrelated random slope with participant. Only significant effects involving Word type and the exposure conditions will be discussed.

For the test task, we conducted two analyses. The first modeled the fixed effects of Exposure, Target, and their interaction on accuracy and RT. Exposure had seven levels: one for each of the six exposure conditions and one for the Test-only condition. This factor was simple-coded with the Test-only condition as the reference level. The three levels of Target were simple-coded with Unrelated as the reference level. Scaled VOT, scaled trial, and the interaction between scaled prime frequency and scaled target frequency were included as covariates. Participant, talker, and the interaction between auditory prime and visual target were included as random intercepts. Target was included as an uncorrelated random slope for participant. If the first analysis revealed significant differences between the Test-only condition and the exposure conditions related to the Target contrasts, we conducted the second analysis focusing only on the exposure conditions; these modeled the fixed effects of Variability, Similarity, Target, and their interactions. The Variability and Similarity coding schemes were the same as in the exposure task analyses. The Target coding scheme, covariates, and random effects were the same as in first set of test task analyses. Significant effects involving

Table 1: Experiment 1 exposure task performance: Mean percent correct (standard error) by Word type for each combination of Similarity and Variability. Contrast presented in log odds rather than probability to reflect unit of analysis.

Similarity	Variability	N	Real words	Pseudowords	Real words vs. Pseudowords
Direct	Invariant	42	80.17 (1.26)	83.94 (1.39)	-0.257
Direct	Variant	38	78.78 (1.26)	85.84 (1.23)	-0.490
Indirect	Invariant	41	77.84 (1.03)	85.10 (1.60)	-0.486
Indirect	Variant	41	79.26 (1.18)	85.64 (1.22)	-0.445
Control	Invariant	49	79.35 (1.09)	86.71 (1.22)	-0.530
Control	Variant	46	75.56 (1.23)	88.67 (1.42)	-0.929

Target and the exposure conditions will be discussed for the first analysis; only significant effects involving the exposure conditions will be discussed for the second analysis to avoid redundancy.

3.2. Results

3.2.1. Exposure

There was a main effect of Word type ($\chi^2(1, N = 2) = 19.75, p < .001$) that reflected overall higher accuracy on pseudowords than on real words ($\beta = -0.56, SE = 0.13, z = -4.44, p < .001$). Summary statistics by Word type and exposure condition are shown in Table 1.

3.2.2. Test: Comparison to the Test-only group

For accuracy, there were main effects of Target ($\chi^2(2, N = 3) = 200.43, p < .001$) and Exposure ($\chi^2(6, N = 7) = 15.81, p = .015$). Regarding Target, accuracy on Unrelated targets was higher than on either Identity targets ($\beta = -3.13, SE = 0.30, z = -10.60, p < .001$) or Competitor targets ($\beta = -4.46, SE = 0.31, z = -14.16, p < .001$). Regarding Exposure, accuracy was higher in the Test-only condition than in the Control-Invariant condition ($\beta = -0.43, SE = 0.20, z = -2.11, p = .035$). In addition, we observed larger differences in accuracy with Indirect-Variant exposure for both the Unrelated-Competitor contrast ($\beta = -2.66, SE = 1.25, z = -2.13, p = .033$) and the Unrelated-Identity contrast ($\beta = -2.45, SE = 1.16, z = -2.10, p = .036$) compared to the Test-only condition. Summary statistics by Target and Exposure are shown in Table 2 and Figure 3.

For RT, we observed a main effect of Target ($\chi^2(2, N = 3) = 504.91, p < .001$). The main effect of Target reflected faster RTs on Unrelated targets than on either Identity targets ($\beta = 15.28, SE = 7.05, z = 2.17, p = .030$)

Table 2: Experiment 1 test task performance: Mean percent correct and mean response time (standard error) by Target for each Exposure group. Accuracy contrasts presented in log odds rather than probability to reflect unit of analysis.

Exposure	N	Identity	Competitor	Unrelated	Unrelated vs. Identity	Unrelated vs. Competitor
Accuracy						
Test-only	37	91.96 (1.12)	78.39 (3.82)	99.04 (0.47)	2.196	3.344
Direct-Variant	38	90.99 (1.02)	72.84 (3.96)	99.76 (0.17)	3.713	5.038
Direct-Invariant	42	90.98 (0.89)	71.36 (3.65)	99.06 (0.32)	2.347	3.746
Indirect-Variant	41	90.34 (1.36)	74.11 (3.57)	99.88 (0.12)	4.522	5.705
Indirect-Invariant	41	91.60 (1.09)	74.63 (3.98)	99.24 (0.30)	2.489	3.798
Control-Variant	46	91.71 (1.24)	73.00 (3.83)	99.50 (0.25)	2.882	4.290
Control-Invariant	49	90.09 (1.16)	71.11 (3.72)	98.24 (0.41)	1.816	3.123
Reaction time						
Test-only	37	625 (19)	793 (30)	607 (15)	-17	-186
Direct-Variant	38	646 (24)	820 (31)	600 (14)	-46	-221
Direct-Invariant	42	632 (24)	821 (36)	622 (18)	-9	-199
Indirect-Variant	41	632 (17)	828 (26)	625 (15)	-7	-203
Indirect-Invariant	41	647 (25)	809 (31)	618 (17)	-30	-191
Control-Variant	46	617 (14)	796 (23)	606 (12)	-10	-189
Control-Invariant	49	653 (18)	866 (25)	644 (14)	-9	-222

or Competitor targets ($\beta = 180.12$, $SE = 8.19$, $z = 21.99$, $p < .001$). In addition, the Test-only condition was associated with overall faster RTs than the Control-Invariant exposure condition ($\beta = 39.50$, $SE = 15.17$, $z = 2.60$, $p = .009$). Summary statistics by Target and Exposure are shown in Table 2 and Figure 3.

3.2.3. Test: Comparison between Variability and Similarity

The accuracy data revealed significant interactions between the Target contrasts and Exposure. Following up this first analysis, we conducted the second analysis focusing on the effects of Variability and Similarity. We observed a main effect of Variability ($\chi^2(1, N = 2) = 9.52$, $p = .002$) that was qualified by an interaction with Target ($\chi^2(2, N = 3) = 9.29$, $p = .010$). The main effect of Variability reflected overall higher accuracy with Variant compared to Invariant exposure ($\beta = 0.57$, $SE = 0.19$, $z = 3.09$, $p = .002$). The Variability-by-Target interaction reflected a larger difference in accuracy for Variant compared to Invariant exposure for both the Unrelated-Identity contrast ($\beta = -1.59$, $SE = 0.52$, $z = -3.04$, $p = .002$) and the Unrelated-Competitor contrast ($\beta = -1.60$, $SE = 0.58$, $z = -2.76$, $p = .006$).

3.3. Discussion

Experiment 1 compared the effects of exposure to voiceless stops (Direct), voiced stops (Indirect), or voiceless fricatives (Control) on adaptation to

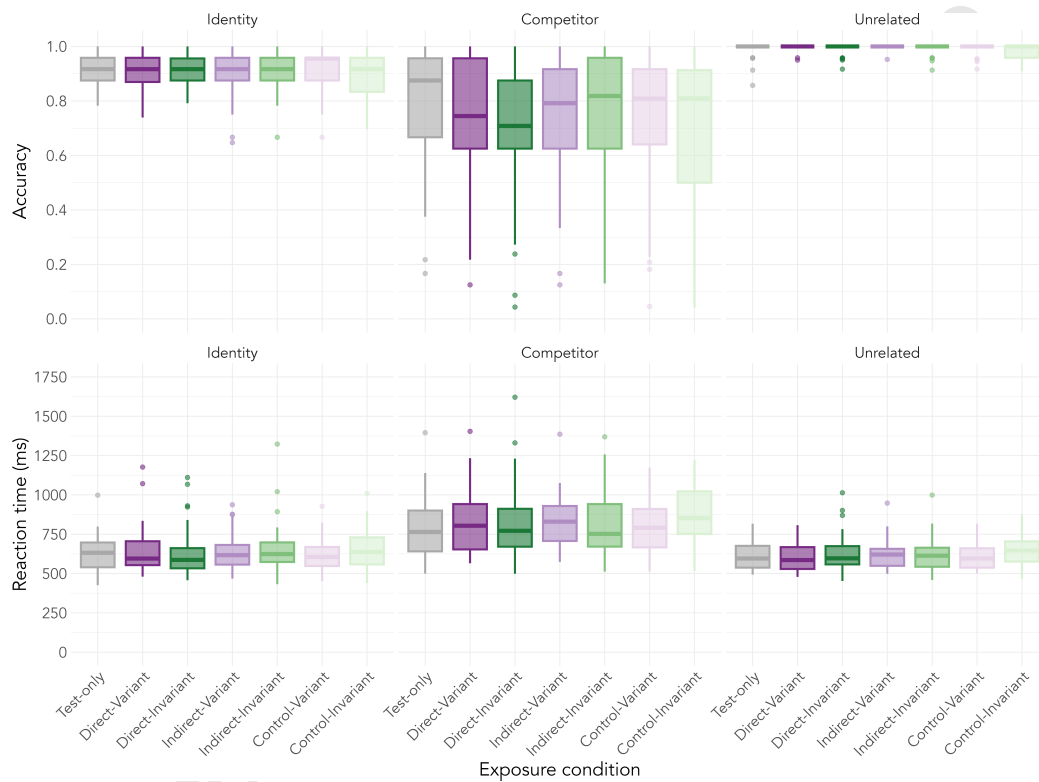


Figure 3: Experiment 1 test task performance. Top row presents boxplots with mean accuracy by participant. Bottom row presents boxplots with mean reaction times for correct responses by participant.

Spanish-accented speech. These three levels of Similarity were crossed with two levels of Variability, such that each talker either produced exemplars of all onsets (Variant) or exemplars of one particular set of onsets (Invariant). Performance on the exposure task reflected a general bias toward labeling tokens as pseudowords rather than as real words; however, listeners were able to distinguish between Spanish-accented real words and pseudowords regardless of exposure condition.

Contrary to our predictions, exposure to Spanish-accented speech did not unequivocally benefit performance on the test task. We observed an increase in the activation of competing targets with Indirect-Variant exposure, with larger differences in accuracy between Competitor and Unrelated targets (*park-bark* vs. *park-wand*) compared to the Test-only condition. We also observed an increase in the activation of matching targets in the Indirect-Variant condition, with larger differences in accuracy between Identity and Unrelated targets (*park-park* vs. *park-wand*) compared to the Test-only condition. Focusing on the unique effects of Variability and Similarity, we found that Variant exposure was associated with increases in both lexical activation and competition compared to Invariant exposure. Overall, our results suggest that listeners responded to Variant exposure by considering more lexical candidates as potential matches to Spanish-accented primes.

This pattern of results contrasts with that of Xie and Myers (2017), which showed that listeners fine-tuned their phonetic categories in response to L2-accented speech. A potential difference in our experimental design was in the construction of the pseudowords. Specifically, we included pseudowords with the experimental onsets that participants were adapting to during exposure. For example, Direct exposure included both the real word *peanut* and the pseudoword **pachine*. While we reasoned that pseudowords like **pachine* should not disrupt perceptual learning of the relevant VOT-/p/ mapping, it is possible that exposure to these mappings in pseudoword contexts disrupted the benefits of exposure to these mappings in disambiguating lexical contexts. To address this issue, we removed these pseudowords in Experiment 2. Instead, participants in the Direct and Indirect groups were both exposed to pseudowords with control onsets like **fachine*. Including control onsets within the Direct and Indirect groups lessened the need for separate Control exposure conditions. To reduce the complexity of the design and home in on the effects of Direct versus Indirect exposure, the Control level of Similarity was not included in Experiment 2.

818 4. Experiment 2: Investigating the specificity of generalization 819 from exposure to Spanish-accented stops

820 4.1. Methods

821 This experiment was the same as Experiment 1, with the exception of
822 the Similarity factor. The Control level was removed, leaving just two levels
823 of Similarity: Direct and Indirect. This change left four between-subjects
824 conditions for the exposure phase: Direct-Variant, Direct-Invariant, Indirect-
825 Variant, and Indirect-Invariant.

826 4.1.1. Participants

827 We recruited 192 participants through Prolific, none of whom had par-
828 ticipated in Experiment 1. All aspects of recruitment and criteria for data
829 exclusion were the same as in Experiment 1. After removing ineligible par-
830 ticipants, 160 remained. After removing participants with poor data quality,
831 154 remained for analysis (Age: $M = 31$, $SD = 6$, $\text{Min} = 18$, $\text{Max} = 40$; Sex:
832 Female = 65, Male = 89; Race: Asian = 5, Black = 17, Multiple selected
833 = 12, Other = 4, White = 116). Data from the participants who completed
834 the experiment without the exposure phase in Experiment 1 were used for
835 comparison.

836 4.1.2. Materials and procedure

837 The talkers, tasks, and procedure were the same as those in Experiment
838 1. Regarding the stimuli, the 288 filler items remained the same. The 72
839 experimental real words for each level of Similarity (Direct and Indirect)
840 also remained the same. The 72 experimental pseudowords, however, were
841 changed. Recall that in Experiment 1, the experimental pseudowords had
842 the same onsets as the experimental real words. It is possible that this
843 design disrupted adaptation to Spanish-accented VOTs by associating them
844 with pseudowords. To improve learning, we replaced the Direct and Indirect
845 experimental pseudowords with the Control experimental pseudowords from
846 Experiment 1. As a result, in Experiment 2, all experimental pseudowords
847 had control onsets.

848 Talker assignment and counterbalancing followed Experiment 1. This
849 resulted in 16 experimental lists for the exposure phase, one for each com-
850 bination of exposure condition (4) and talker assignment (4). All aspects of
851 the test phase remained the same.

Table 3: Experiment 2 exposure task performance: Mean percent correct (standard error) by Word type for each combination of Similarity and Variability. Contrast presented in log odds rather than probability to reflect unit of analysis.

Similarity	Variability	N	Real words	Pseudowords	Real words vs. Pseudowords
Direct	Invariant	36	79.36 (1.44)	87.46 (1.86)	-0.595
Direct	Variant	40	79.89 (1.51)	86.87 (1.84)	-0.511
Indirect	Invariant	39	79.50 (1.30)	88.54 (1.35)	-0.690
Indirect	Variant	39	76.78 (1.21)	90.63 (0.98)	-1.074

4.1.3. Analysis and predictions

The data processing, model fitting, and analysis approaches were the same as in Experiment 1; the only change was implementing sum contrasts for the two levels of Similarity. Prior to analyzing exposure task performance for real words with experimental onsets, we removed responses with RTs less than 50 ms ($N = 22$; 0.10%). We then detected and removed outliers ($N = 995$; 4.49%). Prior to analyzing test task performance for critical prime-target pairs, we removed responses with RTs less than 50 ms ($N = 7$; 0.05%). We then detected and removed outliers ($N = 679$; 4.94%).

4.2. Results

4.2.1. Exposure

There was a main effect of Word type ($\chi^2(1, N = 2) = 23.46, p < .001$) that reflected higher overall accuracy on pseudowords compared to real words ($\beta = -0.94, SE = 0.20, z = -4.84, p < .001$). Summary statistics for each exposure condition are shown in Table 3.

4.2.2. Test: Comparison to the Test-only group

For accuracy, there was a main effect of Target ($\chi^2(2, N = 3) = 190.63, p < .001$), with higher accuracy on Unrelated targets than on either Identity targets ($\beta = -2.52, SE = 0.27, z = -9.42, p < .001$) or Competitor targets ($\beta = -3.85, SE = 0.28, z = -13.81, p < .001$). For RT, there was a main effect of Target ($\chi^2(2, N = 3) = 284.08, p < .001$), which reflected faster RTs on Unrelated targets than on Competitor targets ($\beta = 161.79, SE = 9.76, z = 16.57, p < .001$). Summary statistics by Target and exposure condition are shown in Table 4 and Figure 4.

Table 4: Experiment 2 test task performance: Mean percent correct and mean response time (standard error) by Target for each Exposure group. Accuracy contrasts presented in log odds rather than probability to reflect unit of analysis.

Exposure	N	Identity	Competitor	Unrelated	Unrelated vs. Identity	Unrelated vs. Competitor
Accuracy						
Test-only	37	91.96 (1.12)	78.39 (3.82)	99.04 (0.47)	2.196	3.344
Direct-Variant	40	91.71 (1.26)	72.29 (3.69)	98.67 (0.42)	1.900	3.346
Direct-Invariant	36	90.76 (1.27)	80.65 (2.88)	99.01 (0.39)	2.324	3.181
Indirect-Variant	39	91.10 (1.46)	73.41 (3.57)	99.22 (0.27)	2.525	3.835
Indirect-Invariant	39	91.91 (1.15)	72.97 (3.92)	99.51 (0.24)	2.881	4.318
Reaction time						
Test-only	37	625 (19)	793 (30)	607 (15)	-17	-186
Direct-Variant	40	630 (14)	821 (31)	629 (16)	-1	-192
Direct-Invariant	36	687 (26)	798 (29)	641 (17)	-46	-158
Indirect-Variant	39	614 (19)	827 (43)	610 (19)	-4	-217
Indirect-Invariant	39	653 (26)	793 (25)	628 (19)	-26	-166

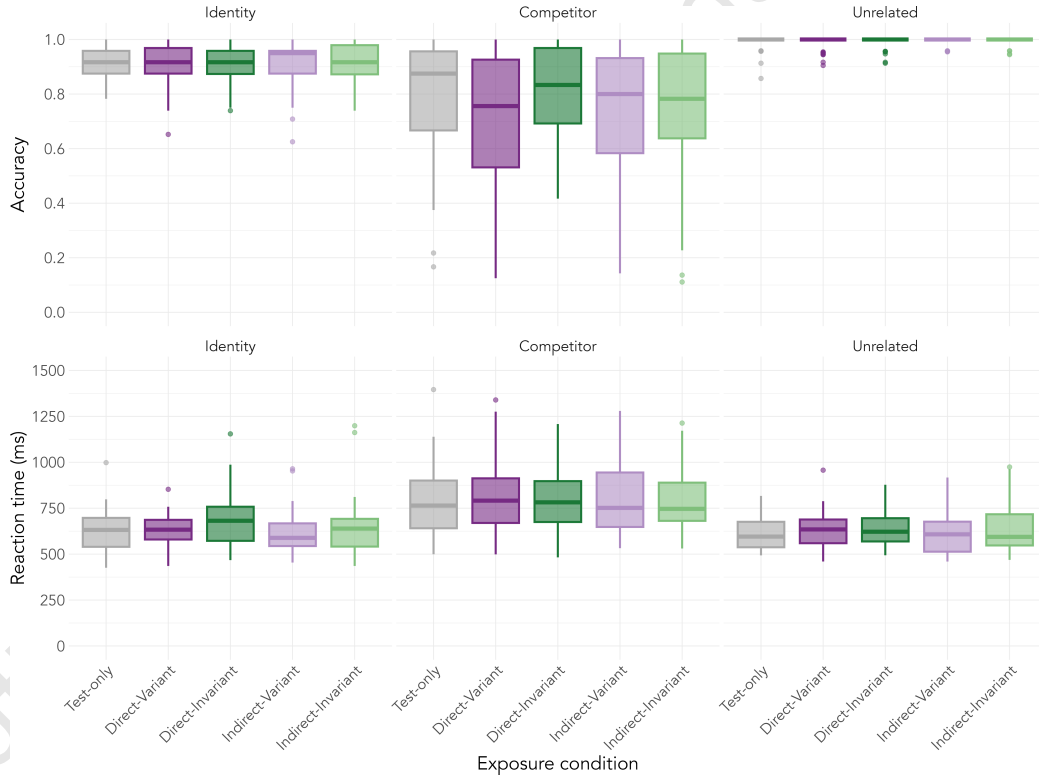


Figure 4: Experiment 2 test task performance. Top row presents boxplots with mean accuracy by participant. Bottom row presents boxplots with mean reaction times for correct responses by participant.

876 4.3. Discussion

877 In Experiment 2, two levels of Similarity—Direct and Indirect—were
878 crossed with two levels of Variability—Variant and Invariant—in order to
879 investigate how exposure to Spanish-accented speech transfers to a novel
880 talker. We provided listeners with even clearer exposure to Spanish-accented
881 stops compared to Experiment 1 by changing the composition of the pseu-
882 dowords. We still observed an overall bias toward labeling tokens as pseu-
883 dowords, but listeners were able to distinguish effectively between the two
884 word types during learning.

885 Removing the pseudowords with experimental onsets from the exposure
886 task also removed the (undesirable) increases in lexical competition that we
887 observed in Experiment 1. In fact, we did not observe any differences in
888 performance between the Test-only and exposure conditions. This lack of an
889 effect of exposure does not align with the wealth of evidence showing that
890 experience with an L2 accent improves perception of a novel L2-accented
891 talker with the same L1 (Bent & Baese-Berk, 2021). Previous studies have
892 generally used either sentence transcription (e.g., Bradlow & Bent, 2008)
893 or primed lexical decision (e.g., Xie & Myers, 2017) to test the strength of
894 adaptation. In Experiments 1 and 2, we used a matching task under the
895 assumption that categorization of short lag VOTs as voiceless stops should
896 change as a function of perception of short lag VOTs as voiceless stops. For
897 example, consider the auditory prime *park* and the visual target *bark*. Par-
898 ticipants needed to decide whether the onset of the token they had heard
899 was a /b/ or not. If they accurately perceived the onset as a /p/, then
900 they would correctly reject *bark* as a match. Thus, accuracy on the match-
901 ing task was the outcome of the categorization process. It is possible that
902 this outcome-based measure was not fine-grained enough to capture subtle
903 changes in perception. To better assess changes in perception, we changed
904 the test task to a cross-modal primed lexical decision task for Experiment 3.

905 5. Experiment 3: Investigating fine-grained perceptual changes 906 from exposure to Spanish-accented stops

907 5.1. Methods

908 The exposure phase was the same as Experiment 2, but the task used
909 in the test phase (cross-modal primed lexical decision task) was different.
910 To better detect subtle changes in perception as a function of exposure to

Spanish-accented speech, we implemented the primed cross-modal lexical decision task from Xie and Myers (2017). We maintained the same three types of Target: Identity, Competitor, and Unrelated. However, participants performed a different task with these targets relative to Experiment 2. Specifically, participants decided whether the visual target was a real English word or not. The auditory primes should increase or decrease RTs on the visual targets as a function of perceptual adaptation to Spanish-accented voiceless stops.

5.1.1. *Participants*

We recruited 195 participants through Prolific; none had participated in Experiments 1 or 2. All aspects of recruitment and criteria for data exclusion were the same as in Experiments 1 and 2. After removing ineligible participants, 170 remained. After removing participants with poor data quality, 158 remained for analysis (Age: $M = 31$, $SD = 6$, Min = 18, Max = 40; Sex: Female = 80, Male = 78; Race: Asian = 4, Black = 20, Multiple selected = 13, Other = 4, White = 117).

Additionally, a new group of 50 participants was also recruited to complete the experiment without the exposure phase. There were 43 participants remaining after checking the eligibility criteria, and 42 participants remained after checking for data quality (Age: $M = 30$, $SD = 5$, Min = 21, Max = 40; Sex: Female = 20, Male = 22; Race: Black = 3, Multiple selected = 5, White = 34).

5.1.2. *Stimuli*

The real words and pseudowords in the exposure task were the same as those in Experiment 2. The auditory primes in the test task were also the same. The only difference was in the visual targets. The unrelated target for each filler prime (144) was replaced with a pseudoword with a different filler onset from the prime. Three additional pseudoword targets were selected for practice. Stimulus selection is described in Appendix C.

5.1.3. *Experimental lists*

The experimental lists for the exposure task were the same as in Experiment 2. Talker assignment was also counterbalanced the same way.

For the critical primes (72), the combinations of auditory prime and visual target were counterbalanced across participants in three experimental lists as in Experiment 2. For the filler primes (144), perfect counterbalancing

946 across these three lists was not possible, since three quarters of the filler
947 items in each list (108) needed to have unrelated pseudoword targets. Three
948 sets of primes, divided evenly by onset, were rotated through the assignment
949 of Identity or Unrelated (pseudoword) target as evenly as possible. This
950 resulted in 12 experimental lists for the test phase, one for each combination
951 of critical prime-target pair (3) and talker assignment (4).

952 5.1.4. *Tasks and procedure*

953 The headphone check, exposure task, and post-experiment questionnaire
954 were the same as Experiment 2. The procedure was also the same. The only
955 difference was in the structure of the test task.

956 The test phase featured the cross-modal primed lexical decision task from
957 Xie and Myers (2017). On each trial, participants first heard a real word (au-
958 ditory prime) and then saw a real word or pseudoword written on the screen
959 (visual target). They indicated whether the visual target was a real English
960 word or not by pressing the *d* or *k* key on their keyboard. The real word
961 response was mapped to the same key as in the exposure task. Participants
962 completed six practice trials followed by 216 main trials presented in random
963 order. Half of the practice trials (3) and half of the main trials (108) required
964 real word responses. All of the trials with critical primes (72) required real
965 word responses.

966 5.1.5. *Analysis*

967 The data processing, model fitting, and analysis approaches were the
968 same as in Experiment 2. Prior to analyzing exposure task performance, we
969 removed responses with RTs less than 50 ms ($N = 10$; 0.04%). We then
970 detected and removed outliers ($N = 1000$; 4.40%). Prior to analyzing test
971 task performance for critical prime-target pairs, we removed responses with
972 RTs less than 50 ms ($N = 53$; 0.37%). We then detected and removed outliers
973 ($N = 1076$; 7.50%).

974 5.2. *Results*

975 5.2.1. *Exposure*

976 There was a main effect of Word type ($\chi^2(1, N = 2) = 33.63, p < .001$)
977 that was qualified by an interaction with Variability ($\chi^2(1, N = 2) = 4.82$,
978 $p = .028$). Pseudoword accuracy was overall higher than real word accuracy
979 ($\beta = -1.13, SE = 0.20, z = -5.80, p < .001$). The interaction with Variability
980 reflected a smaller difference in accuracy between word types with Invariant

Table 5: Experiment 3 exposure task performance: Mean percent correct (standard error) by Word type for each combination of Similarity and Variability. Contrast presented in log odds rather than probability to reflect unit of analysis.

Similarity	Variability	N	Real words	Pseudowords	Real words vs. Pseudowords
Direct	Invariant	41	80.83 (1.18)	86.79 (1.39)	-0.444
Direct	Variant	43	80.03 (1.34)	89.11 (1.40)	-0.714
Indirect	Invariant	39	79.46 (0.99)	88.98 (1.10)	-0.736
Indirect	Variant	35	77.54 (0.95)	90.64 (1.21)	-1.031

981 compared to Variant exposure ($\beta = -0.36$, $SE = 0.17$, $z = -2.20$, $p = .028$).
 982 Summary statistics by Word type and exposure condition are shown in Table
 983 5.

984 5.2.2. Test: Comparison to the Test-only group

985 For accuracy, there was a main effect of Target ($\chi^2(2, N = 3) = 109.15$,
 986 $p < .001$) that was qualified by an interaction with Exposure ($\chi^2(8, N = 9)$
 987 $= 27.25$, $p < .001$). The main effect of Target reflected higher accuracy on
 988 Identity targets ($\beta = 3.28$, $SE = 0.32$, $z = 10.28$, $p < .001$) and Competitor
 989 targets ($\beta = 1.47$, $SE = 0.27$, $z = 5.44$, $p < .001$) compared to Unrelated
 990 targets. The interaction with Exposure was driven by the Indirect-Invariant
 991 condition, which was associated with larger differences in accuracy for both
 992 the Unrelated-Identity contrast ($\beta = 1.61$, $SE = 0.56$, $z = 2.88$, $p = .004$)
 993 and the Unrelated-Competitor contrast ($\beta = 0.84$, $SE = 0.34$, $z = 2.45$, p
 994 $= .014$) compared to the Test-only condition. Summary statistics by Target
 995 and Exposure are shown in Table 6 and Figure 5.

996 For RT, there were main effects of Target ($\chi^2(2, N = 3) = 109.15$, $p <$
 997 $.001$) and Exposure ($\chi^2(4, N = 5) = 37.05$, $p < .001$) that were qualified by
 998 an interaction ($\chi^2(8, N = 9) = 16.54$, $p = .035$). Compared to Unrelated
 999 targets, responses were overall faster for both Identity targets ($\beta = -132.96$,
 1000 $SE = 8.78$, $z = -15.14$, $p < .001$) and Competitor targets ($\beta = -28.02$, $SE =$
 1001 8.60 , $z = -3.26$, $p = .001$). Compared to the Test-only condition, responses
 1002 were overall slower in the all of the exposure conditions: Direct-Variant ($\beta =$
 1003 52.84 , $SE = 13.51$, $z = 3.91$, $p < .001$), Direct-Invariant ($\beta = 81.28$, SE
 1004 $= 13.79$, $z = 5.89$, $p < .001$), Indirect-Variant conditions ($\beta = 54.81$, $SE =$
 1005 14.27 , $z = 3.84$, $p < .001$), and Indirect-Invariant ($\beta = 42.30$, $SE = 13.89$,
 1006 $z = 3.04$, $p = .002$). The Exposure-by-Target interaction was driven by the
 1007 Indirect-Invariant condition, with larger RT differences between Unrelated

Table 6: Experiment 3 test task performance: Mean percent correct and mean response time (standard error) by Target for each Exposure group. Accuracy contrasts presented in log odds rather than probability to reflect unit of analysis.

Exposure	N	Identity	Competitor	Unrelated	Unrelated vs. Identity	Unrelated vs. Competitor
Accuracy						
Test-only	42	96.51 (0.96)	92.40 (1.46)	84.38 (2.69)	-1.633	-0.810
Direct-Variant	43	97.08 (0.66)	90.93 (1.06)	84.74 (2.36)	-1.790	-0.591
Direct-Invariant	41	97.28 (0.69)	89.90 (1.51)	81.67 (2.59)	-2.083	-0.692
Indirect-Variant	35	96.47 (0.71)	92.96 (1.42)	81.42 (3.55)	-1.831	-1.103
Indirect-Invariant	39	98.25 (0.40)	92.65 (1.24)	77.18 (4.03)	-2.812	-1.315
Reaction time						
Test-only	42	583 (12)	674 (14)	731 (26)	148	57
Direct-Variant	43	604 (19)	692 (22)	722 (21)	118	30
Direct-Invariant	41	626 (20)	755 (25)	783 (30)	157	28
Indirect-Variant	35	617 (19)	707 (18)	762 (21)	145	55
Indirect-Invariant	39	589 (14)	708 (17)	749 (21)	160	41

and Identity targets compared to the Test-only condition ($\beta = -41.32$, $SE = 15.91$, $z = -2.60$, $p = .009$). Summary statistics by Target and Exposure are shown in Table 6 and Figure 5.

5.2.3. Test: Comparison between Variability and Similarity

For accuracy, we observed an interaction between Similarity and Target ($\chi^2(2, N = 3) = 13.55$, $p = .001$), which reflected a smaller difference between Unrelated and Competitor targets for Direct compared to Indirect exposure ($\beta = -0.89$, $SE = 0.24$, $z = -3.63$, $p < .001$). In addition, there was a larger difference between Unrelated and Identity targets for Invariant compared to Variant exposure ($\beta = -1.00$, $SE = 0.40$, $z = -2.49$, $p = .013$).

For RT, we observed interactions between Variability and Target ($\chi^2(2, N = 3) = 7.82$, $p = .020$) and between Variability and Similarity ($\chi^2(1, N = 2) = 6.47$, $p = .011$). The Variability-by-Target interaction reflected larger RT differences between Unrelated and Identity targets for Invariant compared to Variant exposure ($\beta = 27.90$, $SE = 11.18$, $z = 2.50$, $p = .013$). The Variability-by-Similarity interaction reflected a larger difference in RTs between Invariant and Variant exposure in the Direct compared to the Indirect conditions ($\beta = -55.23$, $SE = 21.71$, $z = -2.54$, $p = .011$), with the overall fastest RTs in the Direct-Variant condition and the overall slowest in the Direct-Invariant condition. In addition, RT differences between Unrelated and Competitor targets were smaller for Direct compared to Indirect exposure ($\beta = 21.81$, $SE = 10.83$, $z = 2.01$, $p = .044$).

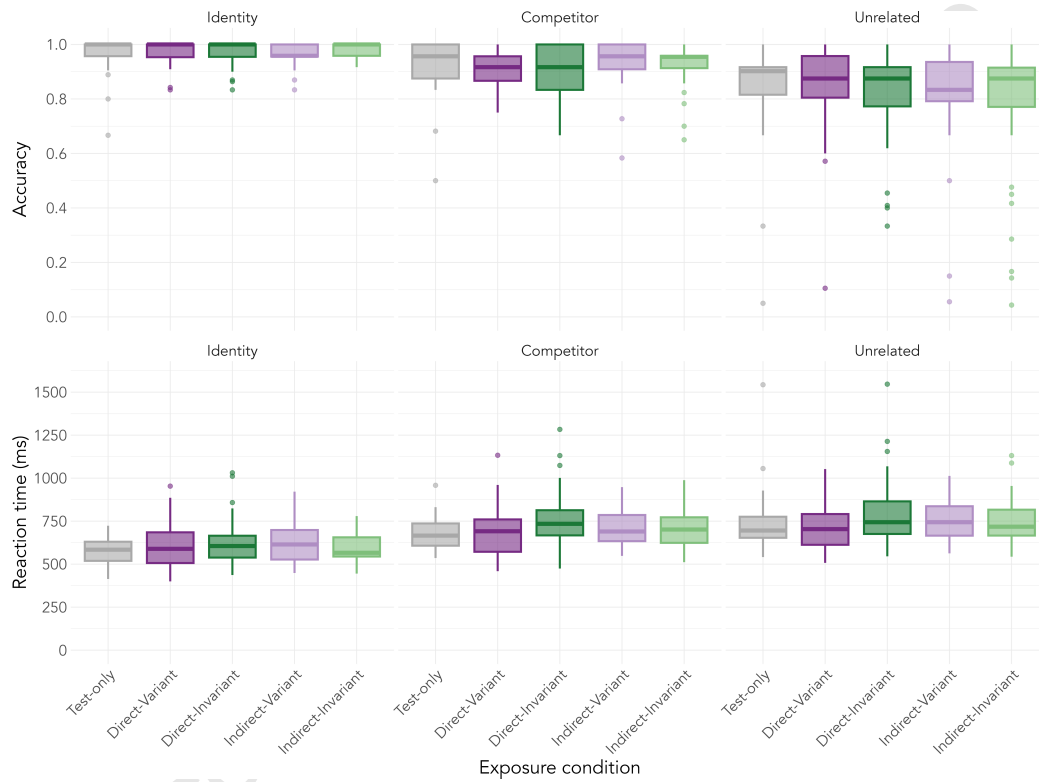


Figure 5: Experiment 3 test task performance. Top row presents boxplots with mean accuracy by participant. Bottom row presents boxplots with mean reaction times for correct responses by participant.

1030 5.3. Discussion

1031 Experiment 3 followed up on the findings of Experiment 2 with a different
1032 test task. In Experiment 2, we used a matching task during the test phase,
1033 which required participants to explicitly compare the visual target to the
1034 auditory prime. We did not observe any differences between the Test-only
1035 condition and the exposure conditions in this task. In Experiment 3, we
1036 took a more implicit approach by using a priming task during the test phase,
1037 which probed the extent to which the auditory primes increased or decreased
1038 activation of the visual targets. In this way, we could investigate changes in
1039 the perception of Spanish-accented English stops.

1040 During exposure, we observed similarly high performance across condi-
1041 tions. Invariant exposure was associated with slightly better performance
1042 than Variant exposure. During test, Indirect-Invariant exposure elicited the
1043 strongest priming effects compared to the Test-only condition. Comparing
1044 the Indirect-Invariant and Test-only conditions, we observed greater identity
1045 priming on both the accuracy and RT measures, indicating an increase in
1046 lexical activation for voiceless stop onsets. We also observed greater com-
1047 petitor priming with Indirect-Invariant exposure, indicating an increase in
1048 lexical competition between voiced and voiceless stops. Together, these re-
1049 sults suggest that Indirect-Invariant exposure led to a relaxation of phonetic
1050 category criteria for voiceless stops.

1051 When we looked at the exposure conditions alone, we saw that Direct
1052 exposure reduced competitor priming relative to Indirect exposure in terms
1053 of both accuracy and RT. We also found that Invariant exposure increased
1054 identity priming relative to Variant exposure on both measures. While the
1055 Direct-Invariant condition itself did not differ from the Test-only condition,
1056 our results suggest that Direct and Invariant exposure are associated with
1057 phonetic recalibration. Consider the auditory prime *park* and the visual tar-
1058 get *bark*, which are minimal pairs that differ only in the voicing of their
1059 onsets. The less the onset of *park* is incorrectly perceived as a /b/, the less
1060 it will activate the competitor *bark*. This means that performance on targets
1061 like *bark* will become more similar to performance on targets like *wand* that
1062 only share a vowel with the prime. We found that Direct exposure induced
1063 this shift in the perception of Spanish-accented VOTs, with Spanish-accented
1064 *park* being perceived less like *bark*. However, this decrease in lexical com-
1065 petition was not accompanied by an increase in lexical activation. Consider
1066 the auditory prime *park* and the visual target *park*; the more the onset of
1067 *park* is correctly perceived as a /p/, the more it will activate the matching

target *park*. This means that performance on targets like *park* will become less similar to performance on targets like *wand* that only share a vowel with the prime. We found that Invariant exposure was associated with this increase in lexical activation, such that Spanish-accented *park* was perceived more like *park*. Overall, Variability and Similarity had differential impacts on lexical activation and competition.

6. General discussion

The current study investigated generalization of exposure to L2-accented speech. We compared two hypotheses with different explanations for how exposure supports talker-independent adaptation. The exposure-to-variability hypothesis argues that increasing covariation between acoustic-phonetic cues during exposure increases performance on a new talker (Baese-Berk et al., 2013; Bradlow & Bent, 2008). The similarity-based hypothesis argues that increasing cue-category overlap between the exposure and test talkers increases performance on the test talker (Xie & Myers, 2017). We contextualized these hypotheses in terms of the ideal adapter framework to investigate their explanatory power within the same theoretical model (Kleinschmidt & Jaeger, 2015). Under this theory, listeners develop mental models of the mappings between acoustic cues (like VOT) and phonetic categories (like /p/, /t/, and /k/) according to informative and useful social (here, L2 accent) and indexical (talker) groupings (Kleinschmidt, 2019). To restate the two competing hypotheses within this framework, exposure to variability enhances the development of robust talker-independent generative models, while exposure-test similarity facilitates the selection of the most appropriate generative models (whether talker-specific or talker-independent). Using a novel experimental approach that improved the scope and precision of how variability and similarity are operationalized, we conducted a series of three experiments. The combined results provide some evidence for the similarity-based hypothesis and other evidence against the exposure-to-variability hypothesis.

The key findings come from Experiment 3. After exposure to three Spanish-accented English talkers, we measured performance on a novel Spanish-accented English talker with a primed cross-modal lexical decision task. The critical items in this task were monosyllabic auditory primes with voiceless stop onsets: for example, *park*, *tune*, and *coal*. Importantly, these items had onset competitors with the same place and manner of articulation: *bark*, *dune*, and *goal*, respectively. In Spanish-accented speech, voiceless stops can

1104 be perceptually confusable with their voiced stop pairs due to differences in
 1105 VOT from L1 English norms. Participants made lexical decisions on the vi-
 1106 sual targets that followed these critical primes. Performance on the different
 1107 target types indexed different aspects of perceptual adaptation. Specifically,
 1108 priming of Identity targets indexed lexical activation of the intended word,
 1109 while priming of Competitor targets indexed lexical activation of perceptu-
 1110 ally similar words. For example, consider the auditory prime *park*. The
 1111 Identity target for this prime is *park*, while the Competitor target for this
 1112 prime is *bark*. If perceptual adaptation to Spanish-accented speech increases
 1113 activation of the lexical item *park* upon hearing *park*, then responses to the
 1114 visual target *park* would increase in accuracy and speed. If perceptual adap-
 1115 tation to Spanish-accented speech decreases activation of the lexical item *bark*
 1116 upon hearing *park*, then responses to the visual target *bark* would decrease
 1117 in accuracy and speed. Ideally, exposure would both increase activation of
 1118 the intended word and decrease activation of competing words. We found
 1119 that Indirect-Invariant exposure increased activation of both voiceless (Iden-
 1120 tity) and voiced (Competitor) targets compared to no exposure. Comparing
 1121 Variability and Similarity directly, we found that Direct exposure was asso-
 1122 ciated with a decrease in lexical competition, while Invariant exposure was
 1123 associated with an increase in lexical activation. To better understand these
 1124 effects, we relate our findings to the exposure-to-variability and similarity-
 1125 based hypotheses.

1126 6.1. Exposure-test similarity

1127 The similarity-based hypothesis comes from Xie and Myers (2017). In
 1128 this study, two groups of participants trained on multiple Mandarin-accented
 1129 talkers. The experimental group was exposed to disambiguating lexical con-
 1130 texts for word-final /d/ (e.g., *overload*), while the other was not. During test
 1131 on a novel Mandarin-accented talker, lexical activation for /d/-final primes
 1132 was higher in the experimental group. Among the Mandarin-accented talkers
 1133 in the training set, one talker’s productions of /d/ were similar to those of
 1134 the test talker on multiple acoustic measures. There was also one talker out
 1135 of the five whose productions of /d/ differed from the test talker’s on these
 1136 metrics. After single-talker exposure to the similar talker, lexical activation
 1137 for /t/-final competitors was lower in the experimental group. Multi-talker
 1138 exposure including the similar talker also facilitated generalization to a novel
 1139 talker. By contrast, single-talker exposure to the dissimilar talker did not
 1140 reduce lexical competition. Overall, the specific way that an L2-accented

1141 talker produces an L2 variety determines the level of generalization to other
1142 talkers. This hypothesis is supported by work on lexically-guided perceptual
1143 retuning showing different patterns of generalization for different phonetic
1144 contrasts (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014).

1145 To operationalize similarity in our study, we exposed participants to dif-
1146 ferent sets of Spanish-accented stops through an auditory lexical decision
1147 task. In Direct conditions, participants gained experience with /p/, /t/, and
1148 /k/ onsets in the context of real words like *peanut*, *terminal*, and *kingdom*,
1149 respectively. Exposure to Spanish-accented voiceless stops trained listen-
1150 ers to shift the VOT distributions for these sound categories from long lag
1151 (English-like) toward short lag (Spanish-like), as illustrated in Figure 1. In
1152 Indirect conditions, participants gained experience with /b/, /d/, and /g/
1153 in the context of real words like *beehive*, *desert*, and *gallop*, respectively. Ex-
1154 posure to Spanish-accented voiced stops trained listeners to shift the VOT
1155 distributions for these sound categories from short lag (English-like) toward
1156 lead (Spanish-like). Shifting the VOT distribution for voiced stops leftward
1157 on the cue continuum had the potential to trigger a more general leftward
1158 shift that would move the VOT distributions for voiceless stops from long
1159 lag to short lag.

1160 Our approach to investigating similarity involved exposure to different
1161 sound categories. By contrast, the approach of Xie and Myers (2017) involved
1162 exposure to different cue distributions over the same sound category. That
1163 study compared the means of three different cues to voicing in word-final
1164 position to determine which talker had the most (dis)similar cue-category
1165 distributions to the test talker; separate groups of participants were then
1166 exposed to either the similar talker or the dissimilar talker. As a result,
1167 their design featured different talkers across exposure conditions. Our design
1168 featured the same talkers across exposure conditions, allowing us to control
1169 the amount of talker-specific acoustic-phonetic variability between levels of
1170 Similarity. In turn, this allowed us to manipulate the constructs of Similarity
1171 and Variability independently. We did not, however, control the exact VOT-
1172 stop distributions that listeners heard during exposure. Rather, we took
1173 advantage of the fact that the VOT distributions for the /p/, /t/, and /k/
1174 onsets in the Direct exposure conditions were not only distinct from those
1175 for the /b/, /d/, and /g/ onsets in the Indirect exposure conditions, but also
1176 similar to those for the critical test items.

1177 Experiment 3 provided some support for the similarity-based hypothesis.
1178 The similarity-based hypothesis predicted a main effect of Similarity, with

1179 better test performance after Direct compared to Indirect exposure. Partic-
 1180 ipants in the Direct condition were also expected to outperform the Test-
 1181 only group. Specifically, Direct exposure was expected to increase identity
 1182 priming and decrease competitor priming. We found that Direct exposure
 1183 decreased lexical competition compared to Indirect exposure consistent with
 1184 phonetic recalibration from exposure to /p/, /t/, and /k/ onsets. However,
 1185 neither of the Direct exposure groups (Direct-Variant or Direct-Invariant)
 1186 differed from the Test-only group. By comparison, Indirect-Invariant ex-
 1187 posure was associated with increased competitor priming compared to the
 1188 Test-only group, as well as increased identity priming. Instead of facilitating
 1189 generalization through fine-tuning phonetic categories for Spanish-accented
 1190 speech, the Indirect-Invariant condition simply increased the number of lexi-
 1191 cal candidates under consideration. These results suggest that exposure-test
 1192 similarity alone is not sufficient for generalization; variability during exposure
 1193 is also a key consideration.

1194 6.2. Exposure to variability

1195 The exposure-to-variability hypothesis is based on the findings of Baese-
 1196 Berk et al. (2013). In that study, exposure to multiple talkers, each with a
 1197 different L2 accent, generalized to a novel talker with an unfamiliar L2 accent.
 1198 Exposure to multiple talkers with the same L2 accent did not generalize to
 1199 an unfamiliar L2 accent (Bradlow & Bent, 2008). However, both types of
 1200 exposure generalized to a novel talker with a familiar L2 accent. The multi-
 1201 accent condition included one Mandarin-accented talker out of five talkers
 1202 with different L1s, while the single-accent condition included five Mandarin-
 1203 accented talkers. In both cases, performance on a novel Mandarin-accented
 1204 talker was higher than in the control condition, suggesting that listeners
 1205 benefit from exposure to systematic covariation across L2 accents. Overall,
 1206 the way that L2 and L1 varieties differ at a high level generalizes across L2
 1207 talkers. This hypothesis is also supported by recent work on cross-accent
 1208 generalization (Bradlow et al., 2023), as well as by the large body of work on
 1209 high variability phonetic training for L2 acquisition (for a review, see Zhang
 1210 et al., 2021).

1211 To operationalize variability in our study, we exposed participants to
 1212 different combinations of talker and onset. In Variant conditions, all three
 1213 exposure talkers produced all three onsets in an experimental group. For
 1214 example, consider the critical onset /p/ in items like *peanut*, *palace*, and *pi-*
 1215 *anist*. Participants in the Direct-Variant exposure group might hear Talker

1216 A produce *peanut*, Talker B produce *palace*, and Talker C produce *pianist*.
1217 As a result, participants would gain experience with each talker’s VOT dis-
1218 tribution for /p/. This condition was designed to encourage the formation
1219 of talker-independent generative models for each onset. By contrast, in In-
1220 variant conditions, each of the three exposure talkers was assigned to one of
1221 the three onsets in an experimental group. This means that participants in
1222 the Direct-Invariant exposure group might only hear Talker A produce items
1223 with /p/ onsets like *peanut*, *palace*, and *pianist*. This condition was designed
1224 to encourage talker-specific generative models for each onset.

1225 Our approach to operationalizing variability allowed us to hold the amount
1226 of exposure to each talker, as well as the talkers themselves, constant be-
1227 tween conditions. By contrast, Bradlow and Bent (2008)’s approach was
1228 to vary the number of talkers encountered during exposure. They found
1229 that multi-talker exposure—comprising five L2-accented talkers with a com-
1230 mon L1—facilitated talker-independent adaptation more than single-talker
1231 exposure—comprising one of the five talkers from the multi-talker condition.
1232 The hypothesis was that exposure to multiple talkers would help listeners
1233 distinguish the common features of an L2 accent from the specific features of
1234 an L2 talker by increasing variation in the input. However, subsequent stud-
1235 ies have shown that the benefit of multi- over single-talker exposure depends
1236 on the specific combination of exposure and test talkers (e.g., Bradlow et
1237 al., 2023). To neutralize any talker-specific effects, we counterbalanced the
1238 assignment of talkers to the exposure and test phases as in Xie et al. (2021)
1239 and Bradlow et al. (2023).

1240 None of the three experiments provided support for the exposure-to-
1241 variability hypothesis. The exposure-to-variability hypothesis predicted a
1242 main effect of Variability, with better test performance after Variant com-
1243 pared to Invariant exposure. Participants in both conditions were also ex-
1244 pected to outperform the Test-only group. In particular, Variant exposure
1245 was expected to increase the activation of intended targets and decrease the
1246 activation of unintended targets. We found that Variant exposure was as-
1247 sociated with increased lexical competition in both Experiments 1 and 3.
1248 This suggests that auditory primes like *park* were perceived more like *bark*
1249 after exposure to variability. In Experiment 1, Variant exposure was also
1250 associated with increased lexical activation, which suggests that the presence
1251 of pseudowords with stop consonant onsets during exposure increased uncer-
1252 tainty about these phonetic categories during test. In fact, when these items
1253 were removed, the effect of Variability on Identity targets (a) disappeared

in Experiment 2 and (b) was reversed in Experiment 3, such that Invariant exposure increased identity priming. Overall, Variant exposure was associated with a relaxation of phonetic category criteria, with listeners considering both /p/ and /b/ onsets for words like *park* (Zheng & Samuel, 2020). The positive association between Invariant exposure and identity priming in Experiment 3 suggests that talker-specific generative models transferred to a novel talker more readily than talker-independent models. Although we included a random intercept for talker in all of our analyses, it is possible that the amount of variability in stop production between our four talkers made talker-independent models less useful for adaptation (see Table B.8 and Figure B.6 in Appendix B). In future work, manipulating between-talker variability by changing the means and shapes of the VOT distributions will provide further insight into the impact of variability on generalization. Including multiple novel talkers with different acoustic properties in the test phase would be another way of investigating the relations between variability and exposure-test similarity in accent adaptation.

6.3. Conclusion

In summary, our results show that talker-independent adaptation to L2-accented speech is facilitated by exposure-test similarity and reducing between-talker variability during exposure. We found that encouraging the development of talker-specific generative models for voiceless stops during exposure reduced competition from voiced stops and increased activation of voiceless stops during test. Our findings are in line with previous work showing that listeners can generalize from exposure to a single talker (Xie et al., 2021; Xie & Myers, 2017). This study is the first to compare variability and similarity while maintaining the same talkers across conditions. This level of control is key to understanding how between-talker variability and talker-specific phonetic detail are used for adaptation. Here, we showed that constraining between-talker variability during exposure allows listeners to use talker-specific phonetic detail during test.

Acknowledgements

This work was supported by the National Science Foundation (BCS-2234907 to Zaharchuk and Van Hell; BCS-2041264 to Van Hell) and the Pennsylvania State University (University Graduate Fellowship to Zaharchuk). The findings and conclusions do not necessarily reflect the views of the funding agencies.

References

- Alexander, J.E.D., & Nygaard, L.C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, 145(6), 3382–3398.
- 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.
- Baigorri, M., Campanelli, L., & Levy, E.S. (2019). Perception of American–English vowels by early and late Spanish–English bilinguals. *Language and Speech*, 62(4), 681–700.
- Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445–459.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bent, T., & Baese-Berk, M. (2021). *Perceptual learning of accented speech* (J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni, Eds.; pp. 428–464). Wiley Blackwell.
- Bieber, R.E., & Gordon-Salant, S. (2022). Semantic context and stimulus variability independently affect rapid adaptation to non-native english speech in young adults. *The Journal of the Acoustical Society of America*, 151(1), 242–255.
- Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (Version 6.1.38). <http://www.praat.org/>
- Bradlow, A.R., Bassard, A.M., & Paller, K.A. (2023). Generalized perceptual adaptation to second-language speech: Variability, similarity, and intelligibility. *The Journal of the Acoustical Society of America*, 154(3), 1601–1613.

- 1319 Bradlow, A.R., & Bent, T. (2008). Perceptual adaptation to non-native
1320 speech. *Cognition*, 106(2), 707–729.
- 1321 Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022).
1322 Does high variability training improve the learning of non-native phoneme
1323 contrasts over low variability training? A replication. *Journal of Mem-*
1324 *ory and Language*, 126, 104352. [https://doi.org/https://doi.org/10.](https://doi.org/https://doi.org/10.1016/j.jml.2022.104352)
1325 1016/j.jml.2022.104352
- 1326 Brysbaert, M., Mandera, P., McCormick, S.F., & Keuleers, E. (2019). Word
1327 prevalence norms for 62,000 English lemmas. *Behavior Research Methods*,
1328 51, 467–479.
- 1329 Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence
1330 from 18 languages. *Journal of Phonetics*, 27(2), 207–229.
- 1331 Chodroff, E., Golden, A., & Wilson, C. (2019). Covariation of stop voice on-
1332 set time across languages: Evidence for a universal constraint on phonetic
1333 realization. *The Journal of the Acoustical Society of America*, 145(1),
1334 EL109–EL115.
- 1335 Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic
1336 realization: Covariation of stop consonant VOT in American English.
1337 *Journal of Phonetics*, 61, 30–47.
- 1338 Choi, J.Y., & Perrachione, T.K. (2019). Time and information in perceptual
1339 adaptation to speech. *Cognition*, 192, 103982.
- 1340 Clarke, C.M., & Garrett, M.F. (2004). Rapid adaptation to foreign-accented
1341 English. *The Journal of the Acoustical Society of America*, 116(6), 3647–
1342 3658.
- 1343 Clayards, M. (2017). Individual talker and token covariation in the produc-
1344 tion of multiple cues to stop voicing. *Phonetica*, 75(1), 1–23.
- 1345 Clayards, M., Tanenhaus, M.K., Aslin, R.N., & Jacobs, R.A. (2008). Percep-
1346 tion of speech reflects optimal use of probabilistic speech cues. *Cognition*,
1347 108(3), 804–809.
- 1348 Ethnologue. (2024). *What are the top 200 most spoken languages?* [https:](https://www.ethnologue.com/insights/ethnologue200/)
1349 [//www.ethnologue.com/insights/ethnologue200/](https://www.ethnologue.com/insights/ethnologue200/)
- 1350 Flege, J.E., & Bohn, O. (2021). The revised speech learning model (SLM-r).
1351 *Second Language Speech Learning: Theoretical and Empirical Progress*,
1352 3–83.
- 1353 Flege, J.E., & Eefting, W. (1987). Production and perception of english
1354 stops by native spanish speakers. *Journal of Phonetics*, 15(1), 67–83.
- 1355 Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*
1356 (Third). Sage. <https://socialsciences.mcmaster.ca/jfox/Books/>

- Companion/
- Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52(12), 5186–5201.
- 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.
- 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.
- 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.
- Lisker, L., & Abramson, A.S. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences*, 563–567.
- Lisker, L., & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3), 384–422.
- Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6, 1171.
- Nagle, C.L., & Baese-Berk, M.M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, 44(2), 580–605.
- Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J.K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203.
- R Core Team. (2022). *R: A language and environment for statistical computing*. 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.
- Sidaras, S.K., Alexander, J.E.D., & 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.

- 1395 Tzeng, C.Y., Alexander, J.E.D., Sidaras, S.K., & Nygaard, L.C. (2016). The
1396 role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance*,
1397 42(11), 1793.
1398
- 1399 Tzeng, C.Y., Russell, M.L., & Nygaard, L.C. (2024). Attention modulates
1400 perceptual learning of non-native-accented speech. *Attention, Perception, & Psychophysics*, 86(1), 339–353.
1401
- 1402 United Nations. (2024). *Standard country or area codes for statistical use*
1403 (M49). <https://unstats.un.org/unsd/methodology/m49/>
1404
- 1404 Vicente, M.L.C. (1986). El VOT de las oclusivas sordas y sonoras españolas.
1405 *Estudios de Fonética Experimental*, 91–110.
- 1406 Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in
1407 the perceptual learning of non-native-accented speech sounds. *Phonetica*,
1408 64(2-3), 122–144.
- 1409 Williams, L. (1977). The voicing contrast in Spanish. *Journal of Phonetics*,
1410 5(2), 169–184.
- 1411 Witteman, M.J., Weber, A., & McQueen, J.M. (2013). Foreign accent
1412 strength and listener familiarity with an accent codetermine speed of
1413 perceptual adaptation. *Attention, Perception, & Psychophysics*, 75(3),
1414 537–556.
- 1415 Woods, K.J., Siegel, M.H., Traer, J., & McDermott, J.H. (2017). Head-
1416 phone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79, 2064–2072.
1417
- 1418 Xie, X., & Jaeger, T.F. (2020). Comparing non-native and native speech:
1419 Are L2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5), 3322–3347.
1420
- 1421 Xie, X., Liu, L., & Jaeger, T.F. (2021). Cross-talker generalization in the
1422 perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*.
1423
- 1424 Xie, X., & Myers, E.B. (2017). Learning a talker or learning an accent:
1425 Acoustic similarity constrains generalization of foreign accent adaptation
1426 to new talkers. *Journal of Memory and Language*, 97, 30–46.
- 1427 Xie, X., Theodore, R.M., & Myers, E.B. (2017). More than a boundary shift:
1428 Perceptual adaptation to foreign-accented speech reshapes the internal
1429 structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 43(1), 206–217.
1430
- 1431 Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., &
1432 Jaeger, T.F. (2018). Rapid adaptation to foreign-accented speech and its

- 1433 transfer to an unfamiliar talker. *The Journal of the Acoustical Society of*
1434 *America*, 143(4), 2013–2031.
- 1435 Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability
1436 in nonnative phonetic learning: A systematic review and meta-analysis.
1437 *Journal of Speech, Language, and Hearing Research*, 64(12), 4802–4825.
- 1438 Zheng, Y., & Samuel, A.G. (2020). The relationship between phonemic cat-
1439 egory boundary changes and perceptual adjustments to natural accents.
1440 *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
1441 46(7), 1270.

Appendix A. Norming study

Appendix A.1. Participants

Prior to conducting Experiments 1-3, we recruited a separate group of participants ($N = 688$) from the Penn State subject pool to norm the auditory stimuli. Participants provided implied consent in line with Penn State IRB policies and were compensated 0.5 class credits after completing the study, which took 15-30 minutes. Eligible participants were between the ages of 18 and 40 years, spoke English as their first and only fluent language, had normal hearing, had normal or corrected-to-normal vision, and did not have a history of language-related disorders. We removed ineligible participants ($N = 34$) and those with poor data quality ($N = 83$) from further analysis. This left 571 participants (demographic information was not collected).

Appendix A.2. Stimuli

Stimulus selection began by downloading real words with one to four syllables, three to eight letters, and one to two morphemes from the English Lexicon Project (ELP) restricted lexicon (Balota et al., 2007). Within this set of items, we limited our search to words with experimental or filler onsets followed directly by a vowel (e.g., *peach* was considered but *preach* was not). We removed duplicate word stems with different suffixes (e.g., *paints* and *painting* were removed but *paint* was kept). We also removed any inappropriate, harmful, or distracting words and word stems. The remaining set of items will be referred to as the ELP pool.

Multisyllabic stimulus selection began by drawing real words with two to four syllables, five to eight letters, and one to two morphemes from the ELP pool. We removed words with minimal pairs between the critical and competitor groups (e.g., *pocket* and *docket* were removed but *socket* and *loket* were kept). Once the set of options was established, real words were selected and pseudowords were created. Pseudowords were created by changing the onsets of real words with filler onsets that had not been selected as potential real words for the study. For experimental pseudowords, onsets were assigned according to the cross-condition onset groups (e.g., *machine* became **pachine*, **bachine*, and **fachine*). For filler pseudowords, one of the other five filler onsets was substituted for the existing filler onset (e.g., *medicine* became **hedicine*). In total, 544 multisyllabic real words and 555 multisyllabic pseudowords were normed.

1477 Monosyllabic stimulus selection began by drawing real words with one
1478 syllable, three to six letters, and one morpheme from the ELP pool. From
1479 this set of options, we pulled all words with critical onsets that had cross-
1480 experimental minimal pairs with competitor onsets (e.g., *park-bark*). Both
1481 members of each critical-competitor minimal pair were normed. We also
1482 selected words with filler onsets that had a minimal pair with a different filler
1483 onset (e.g., *mall-hall*). In total, 319 monosyllabic real words were normed.

1484 Norming took place in three waves of testing. Overall, there were nine
1485 experimental lists of 180 to 540 items: three in the first wave, two in the
1486 second, and four in the third. Lists always contained an equal number of
1487 real words and pseudowords, and items from each experimental onset groups
1488 were always presented in separate lists.

1489 *Appendix A.3. Recording*

1490 One female L1-accented English talker from the US (Talker 0; the first
1491 author) recorded all items for the norming study. Audio was captured with a
1492 head-worn condenser microphone (Shure SM35-XLR) connected to an audio
1493 interface (Sound Devices USBPre2) and recorded with Praat (Boersma &
1494 Weenink, 2021) in mono at 44.1 kHz in a sound-attenuated booth. After
1495 annotating the recordings and extracting the individual sound files, stimuli
1496 were normalized to 70 dB and had 50 ms of silence added to the beginning
1497 and end.

1498 *Appendix A.4. Task and procedure*

1499 The norming study was conducted online using Pavlovia. At the beginning
1500 of the study, participants responded to a yes/no question for each eligibility
1501 criterion. Ineligible participants were not able to complete the study.

1502 The study featured an auditory lexical decision task. Participants were
1503 randomly assigned to an experimental list. On each trial, participants in-
1504 dicated whether an auditory stimulus was a real English word or not by
1505 pressing the *d* or *k* key on their keyboard. One of the two response-key
1506 relations—real-*d* or real-*k*—was assigned randomly to each participant. The
1507 total number of trials varied by list.

1508 *Appendix A.5. Analysis*

1509 Analyses for the norming study were conducted with R version 4.5.1 using
1510 the *stats* package (R Core Team, 2022). Each wave of testing was analyzed

Table B.7: Talker background information.

Talker	Region	Age of English acquisition	Age of arrival in US	Age at time of recording
1	Veracruz, Mexico	10	18	26
2	Quintana Roo, Mexico	13	26	28
3	Puebla, Mexico	5	24	27
4	Mexico City, Mexico	5	28	31

separately. The experimental lists within each wave were combined for analysis. We first conducted one-sample t-tests comparing each participant’s accuracy to chance (50%) to check for data quality. Data from participants whose performance was indistinguishable from chance were removed from further analysis. Next, we conducted one-sample t-tests comparing the accuracy on each item to chance (50%). Items that did not yield above-chance accuracy were removed from further consideration. All three variants of each experimental pseudoword and both members of each cross-experimental minimal pair needed to have above-chance accuracy in order to be considered for final selection. The set of items that passed the norming phase will be referred to as the selection pool. The final set of stimuli for Experiments 1-3 was chosen from the selection pool as described in Section 3.1.3.

Appendix B. Spanish-accented talker details

Appendix B.1. Talker selection and recording

Seven female Spanish-English bilinguals from Latin America and Spain were recruited to be talkers for the experiment. Talkers were compensated with one \$20 Amazon giftcard per recording session (30-60 minutes each; 1-2 total). Recording procedures followed those in the norming study. We used the recordings from four out of the seven talkers as experimental stimuli (Talkers 1-4). Language background information for each talker is provided in Table B.7.

Appendix B.2. Stop VOT distributions

Descriptive statistics for each talker’s real word VOTs by onset are provided in Table B.8.

Figure B.6 visualizes the cue-category distributions for each talker and voiceless onset with respect to the L1 US English mean.

Table B.8: Mean, standard deviation, and range of VOTs by onset by talker.

Phase	Onset	<i>N</i>	Talker 1			Talker 2			Talker 3			Talker 4		
			<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range
Exposure	/b/	24	-21	51	-132-48	15	27	-87-77	-44	52	-150-31	-69	53	-133-103
	/d/	24	2	42	-100-39	21	22	-75-53	8	29	-54-33	-78	53	-177-44
	/g/	24	2	55	-104-68	31	25	-75-51	-5	51	-148-47	-64	65	-229-67
	/p/	24	73	26	24-122	61	41	11-160	82	26	37-123	21	17	6-69
	/t/	24	72	20	36-112	65	22	25-109	91	28	46-157	41	15	18-74
	/k/	24	82	28	33-151	88	21	49-123	96	23	43-133	48	16	25-77
Test	/p/	24	71	28	19-126	60	35	10-118	115	39	15-201	26	24	8-106
	/t/	24	78	21	27-109	61	20	28-106	102	23	60-142	53	27	8-99
	/k/	24	94	21	70-171	94	20	50-133	119	40	65-244	65	30	21-124

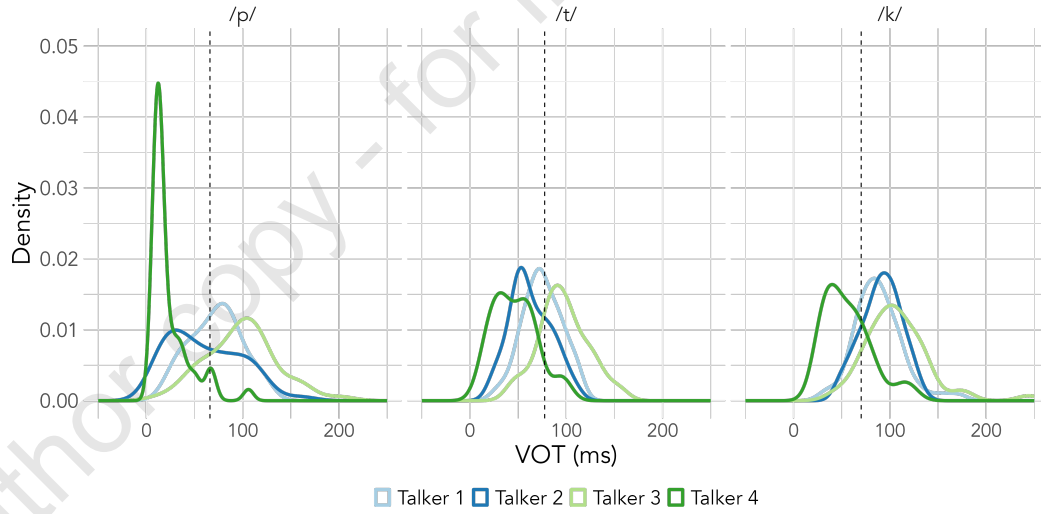


Figure B.6: Smoothed density estimates of VOT by talker and onset. Dotted lines represent L1 US English means from Chodroff and Wilson (2019).

Appendix C. Stimulus selection

Appendix C.1. Experiment 1

The exposure task featured multisyllabic real words and pseudowords with experimental and filler onsets from the selection pool. The test task featured two types of stimuli: auditory primes and visual targets. Auditory primes were monosyllabic real words with critical or filler onsets from the selection pool. Visual targets in the test task were monosyllabic real words with critical, competitor, or filler onsets.

Each auditory prime had three visual targets: one was the prime itself, one was its minimal pair, and one was unrelated. The first two were determined by the prime, but the third needed to be selected separately. Each unrelated target had a filler onset and was chosen to have the same vowel as the prime but a different onset and offset. These items were primarily from the selection pool, but some were from the larger ELP pool.

During auditory stimulus selection, we considered a number of lexical factors. From the ELP, we included word length, orthographic neighborhood density (OND), phonological neighborhood density (PND), and US Zipf frequency. From Brysbaert et al. (2019), we included percent known and prevalence. We also included the position of lexical stress, calculated from the pronunciation information provided by the ELP, and mean lexical decision accuracy from norming (LDT). The final set of exposure real words was chosen such that the four groups of stimuli—Direct, Indirect, Control, and Filler—were matched on each of these parameters. We conducted two-sample t-tests comparing each group on each parameter to ensure that they were not significantly different ($ps > .05$). The final set of auditory primes for the test task was selected in a similar way. We conducted two-sample t-tests comparing the Critical and Filler primes on every parameter except for lexical stress, which was not relevant for monosyllabic words ($ps > .05$). We also conducted paired t-tests between Critical primes and their Competitor pairs ($ps > .05$). The final set of exposure pseudowords was chosen such that the four groups of stimuli were matched on length, position of lexical stress, and mean lexical decision accuracy. The other parameters were either not available (OND and PND) or not relevant (US Zipf frequency, percent known, prevalence) for the pseudowords. The relevant parameters for each group of stimuli are shown in Table C.9.

Table C.9: Mean and standard deviation of stimulus parameters by condition and task.

Phase	Word type	Condition	N	Length	LDT	Lexical stress	OND	PND	Frequency	Percent known	Prevalence
Exposure	Real word	Direct	72	6.85 (0.93)	0.89 (0.09)	1.11 (0.32)	0.56 (1.40)	1.12 (2.13)	3.38 (0.74)	0.98 (0.02)	2.12 (0.29)
		Indirect	72	6.82 (0.98)	0.89 (0.08)	1.21 (0.41)	0.62 (1.34)	1.14 (2.04)	3.45 (0.72)	0.99 (0.02)	2.16 (0.29)
		Control	72	6.93 (0.95)	0.87 (0.09)	1.15 (0.36)	0.67 (1.17)	1.76 (2.76)	3.39 (0.77)	0.98 (0.03)	2.11 (0.38)
	Pseudoword	Filler	144	6.83 (0.95)	0.87 (0.10)	1.12 (0.32)	0.78 (1.52)	1.49 (2.66)	3.42 (0.77)	0.98 (0.03)	2.14 (0.35)
		Direct	72	7.08 (0.87)	0.89 (0.06)	1.22 (0.42)					
		Indirect	72	7.08 (0.87)	0.88 (0.07)	1.22 (0.42)					
		Control	72	7.08 (0.87)	0.88 (0.06)	1.22 (0.42)					
		Filler	144	6.93 (1.01)	0.88 (0.05)	1.14 (0.35)					
	Real word	Critical prime	72	3.88 (0.63)	0.88 (0.07)		11.36 (5.21)	21.85 (7.96)	4.15 (0.94)	0.99 (0.04)	2.22 (0.34)
		Competitor pair	72	3.90 (0.65)	0.88 (0.08)		10.39 (5.02)	21.15 (7.12)	4.09 (1.11)	0.98 (0.05)	2.10 (0.42)
		Filler prime	144	3.90 (0.58)	0.90 (0.08)		10.49 (4.70)	22.27 (8.89)	4.15 (0.97)	0.99 (0.02)	2.22 (0.32)

Competitor pairs were presented as visual targets only

Appendix C.2. Experiment 2

No new stimuli were selected for this experiment.

Appendix C.3. Experiment 3

Potential pseudowords were downloaded from the ELP's set of normed pseudowords. The best possible match was selected for each prime by (orthographic) vowel and number of letters.

Appendix D. Post-experiment questionnaire

Appendix D.1. Questionnaire items

Here we describe the items in the questionnaire and how they were coded for analysis. Items requiring ratings were collected on five-point scales anchored at the endpoints with labels containing the modifier “very” and a relevant adjective (e.g., “very weak” and “very strong” for accent strength). Items requiring categorical responses were collected by presenting a relevant set of options to choose from (see below). The options for these items were presented in alphabetical order.

Appendix D.1.1. Talker questionnaire items

Participants indicated whether the talker had an accent or not, then rated the strength of the accent (Talker accent strength) and identified the type of accent. For the type of accent, participants selected all that applied from the following options: city or region in the US; city or region outside the US; social, racial, or ethnic group; another language; and speech or language impairment.

Next, participants rated how well they understood the talker (Talker comprehensibility) and how easy it was to understand the talker (Talker ease).

They also identified the talker’s L1 and proficiency in their L1. Participants selected from a list of the top 10 most spoken languages in the world (Ethnologue, 2024), as well as a “None of these” option. Participants then indicated whether the talker knew a second language or not; if so, they identified the talker’s L2 and rated their proficiency. The same options for identifying the talker’s L1 were provided for the L2.

Regarding the L1 and L2 proficiency ratings, we included whichever rating was provided for English (Talker English rating). Regarding the first and second languages, we collapsed across the two responses and coded them according to typological distance from Spanish (Talker language accuracy). Responses that only included English were coded as 0. Otherwise, responses identifying either the L1 or L2 as Spanish were coded as 5, Portuguese as 4, French as 3, another Indo-European language (Urdu, Hindi, Bengali, or Russian) as 2, and a language from a different family (Mandarin, Standard Arabic, or Other) as 1.

Finally, participants identified the talker’s place of origin. They began by selecting a global region—Africa, Americas, Asia, Europe, or Oceania—followed by a sub-region within the chosen region (United Nations, 2024). If the Americas region was selected, participants were also able to select a specific country within their chosen sub-region. We coded responses by distance from the correct sub-region (Talker region accuracy), such that Americas-Central was 5, Americas-Northern was 4, Americas-Caribbean was 3, Americas-Southern was 2, Europe was 1, and all other responses were 0.

Appendix D.1.2. Participant questionnaire items

Participants indicated whether they had an accent or not, then rated the strength of the accent (Own accent strength) and identified the type of accent from the options described above. To measure the overlap between a participant’s perception of their own accent and their perception of the talker’s accent, we divided the number of labels in common by the total number of unique labels from both responses (Own accent comparison). For example, if a participant selected “social, racial, or ethnic group” for their own accent and “social, racial, or ethnic group” and “another language” for the talker’s accent, their value would be 0.5 (i.e., 1 common label / 2 total unique labels).

In addition, participants identified their L1 and rated their proficiency. They also indicated whether they knew a second language or not; if they responded “yes”, they identified the L2 from the options described above

Table D.10: Experiment 1 participant and talker questionnaire results: Mean response (standard deviation) [number of responses] for each exposure group.

Measure	Direct-Variant	Direct-Invariant	Indirect-Variant	Indirect-Invariant	Control-Variant	Control-Invariant
Talker ease	2.50 (0.65) [38]	2.67 (0.72) [42]	2.73 (0.81) [41]	2.85 (0.73) [41]	2.59 (0.72) [46]	2.94 (0.83) [49]
Talker comprehensibility	3.05 (0.73) [38]	2.98 (0.84) [42]	3.20 (0.75) [41]	3.12 (0.71) [41]	2.93 (0.61) [46]	3.20 (0.71) [49]
Talker accent strength	4.16 (0.79) [38]	4.00 (0.80) [42]	4.15 (0.66) [40]	3.93 (0.75) [41]	4.00 (0.82) [46]	3.62 (0.87) [48]
Talker English rating	3.30 (0.95) [30]	3.21 (0.92) [28]	3.26 (0.74) [35]	3.23 (0.88) [35]	3.39 (0.75) [38]	3.57 (0.89) [46]
Talker language accuracy	2.32 (1.63) [38]	2.50 (1.58) [42]	2.68 (1.59) [41]	2.29 (1.33) [41]	2.65 (1.58) [46]	2.61 (1.64) [49]
Talker region accuracy	1.08 (1.42) [38]	1.17 (1.54) [42]	1.12 (1.47) [41]	0.68 (0.85) [41]	1.09 (1.60) [46]	1.12 (1.56) [49]
Own accent comparison	0.04 (0.18) [38]	0.00 (0.00) [42]	0.00 (0.00) [41]	0.03 (0.16) [41]	0.00 (0.00) [46]	0.01 (0.05) [48]
Own accent strength	0.66 (1.32) [38]	0.57 (1.11) [42]	1.02 (1.59) [41]	0.98 (1.52) [41]	0.96 (1.48) [46]	0.55 (1.06) [49]

Table D.11: Experiment 2 participant and talker questionnaire results: Mean response (standard deviation) [number of responses] for each exposure group. The Test-only group is included in the table for Experiment 1.

Measure	Direct-Variant	Direct-Invariant	Indirect-Variant	Indirect-Invariant
Talker ease	2.85 (0.74) [40]	2.86 (0.80) [36]	2.56 (0.72) [39]	2.72 (0.86) [39]
Talker comprehensibility	3.15 (0.58) [40]	2.94 (0.89) [36]	2.87 (0.77) [39]	2.95 (0.86) [39]
Talker accent strength	3.90 (0.78) [40]	4.06 (0.68) [35]	3.95 (0.72) [39]	3.97 (0.82) [38]
Talker English rating	3.48 (0.71) [33]	3.31 (0.82) [32]	3.57 (0.86) [30]	3.40 (0.81) [35]
Talker language accuracy	2.20 (1.32) [40]	2.11 (1.14) [36]	2.49 (1.50) [39]	2.85 (1.71) [39]
Talker region accuracy	0.65 (0.95) [40]	0.72 (1.16) [36]	0.95 (1.38) [39]	1.18 (1.64) [39]
Own accent comparison	0.02 (0.11) [40]	0.00 (0.00) [35]	0.00 (0.00) [39]	0.01 (0.04) [39]
Own accent strength	0.47 (1.06) [40]	0.92 (1.52) [36]	0.59 (1.31) [39]	0.82 (1.62) [39]

and rated their proficiency. Participants also identified their place of origin as described above and indicated whether they currently lived in this location or not. At the end of the questionnaire, participants provided their age and indicated whether they had normal/corrected-to-normal vision or not, whether they had normal hearing or not, and whether they had a history of language disorders or not.

Appendix D.2. Participant and talker results

Descriptive statistics for each group of participants are provided by experiment in Tables D.10, D.11, and D.12. In addition, all participants reported that their L1 was English or did not provide a response (*NA* responses = 6 across experiments). Only 10 participants across all three experiments reported knowledge of a second language; these were French (2), Portuguese (1), or Other (7). Basic demographics are reported in their respective experiment sections in the main manuscript.

Table D.12: Experiment 3 participant and talker questionnaire results: Mean response (standard deviation) [number of responses] for each exposure group.

Measure	Test-only	Direct-Variant	Direct-Invariant	Indirect-Variant	Indirect-Invariant
Talker ease	3.05 (0.96) [42]	2.51 (0.80) [43]	2.88 (0.75) [41]	2.83 (0.75) [35]	2.82 (0.76) [39]
Talker comprehensibility	3.29 (0.92) [42]	2.77 (0.78) [43]	3.07 (0.72) [41]	3.00 (0.80) [35]	3.08 (0.77) [39]
Talker accent strength	3.55 (0.83) [38]	3.98 (0.75) [42]	3.98 (0.69) [41]	3.97 (0.82) [35]	3.82 (0.85) [39]
Talker English rating	3.47 (0.92) [38]	3.38 (0.79) [37]	3.58 (0.92) [31]	3.36 (0.87) [28]	3.46 (0.85) [35]
Talker language accuracy	2.14 (1.42) [42]	2.49 (1.40) [43]	2.73 (1.45) [41]	2.34 (1.61) [35]	2.33 (1.46) [39]
Talker region accuracy	1.02 (1.37) [42]	1.12 (1.64) [43]	1.10 (1.41) [41]	0.80 (0.93) [35]	1.05 (1.50) [39]
Own accent comparison	0.00 (0.00) [38]	0.02 (0.08) [42]	0.00 (0.00) [41]	0.01 (0.04) [35]	0.00 (0.00) [39]
Own accent strength	0.48 (1.06) [42]	0.74 (1.26) [43]	0.88 (1.36) [41]	0.94 (1.47) [35]	0.92 (1.53) [39]