1   Unravelling the time-course of listener adaptation to an unfamiliar talker

2   Maryann Tan[1,2] & T. Florian Jaeger[2,3]

3   [1] Centre for Research on Bilingualism, University of Stockholm

4   [2] Brain and Cognitive Sciences, University of Rochester

5   [3] Computer Science, University of Rochester

6   Author Note

Abstract

YOUR ABSTRACT GOES HERE. All data and code for this study are shared via OSF, including the R markdown document that this article is generated from, and an R library that implements the models we present.

Keywords: speech perception; adaptation; incremental changes; distributional learning

Word count: X

Unravelling the time-course of listener adaptation to an unfamiliar talker

TO-DO

0.1 Highest priority

- MARYANN

0.1.1 Lower Priority

- Decide on PSE vs. category boundary
- standardize BE vs. AE spelling (categoriz/sation, label(l)ed, synthesiz/sed etc.)

0.2 To do later

- Everyone: Eat ice-cream and perhaps have a beer.

1 Introduction

One of the hallmarks of human speech perception is its adaptivity. Listeners' interpretation of acoustic input can change within minutes of exposure to an unfamiliar talker, supporting robust speech recognition across talkers (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Xie, Liu, & Jaeger, 2021; Xie et al., 2018). Recent reviews have identified distributional learning of marginal cue statistics ('normalization,' Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011) or the statistics of cue-to-category mappings as an important mechanisms affording this adaptivity ('representational learning,' Clayards, Tanenhaus, Aslin, & Jacobs, 2008; D. F. Kleinschmidt & Jaeger, 2015; **idemaru-hold2011?**; **davis-sohoglu2020?**; for review, **schertz-clare2019?**; **xie2023?**). This hypothesis has gained considerable influence over the past decade, with findings that changes in listener perception are qualitatively predicted by the statistics of exposure stimuli (Bejjanki, Clayards, Knill, & Aslin, 2011; Clayards et al., 2008; Nixon, Rij, Mok, Baayen, & Chen, 2016; Tan, Xie, & Jaeger, 2021; **idemaru2021?**; **kleinschmidt2012?**; **munson2011-thesis?**; **theodore2019distributional?**; for important caveats, see **harmon2018?**).

We investigate an important constraints on this type of adaptivity that is suggested by recent findings. D. F. Kleinschmidt and Jaeger (2016) exposed L1-US English listeners to recordings of /b/-/p/ minimal pair words like *beach* and *peach* that were acoustically manipulated. Separate groups of listeners were exposed to distributions of voice onset times (VOTs)—the primary cue distinguishing between *beach* and *peach*—that were shifted by XXX to XXX msecs, respectively, relative to what one might expect from a 'typical' talker (Figure 1A). In line with the distributional learning hypothesis, listeners' category boundary or point of subjective equality (PSE)—i.e., the VOT for which listeners are equally likely to respond "d" or "t"—shifted in the same direction as the exposure distribution (Figure 1B). Also in line with the distributional learning hypothesis, these shifts were larger the further the exposure distributions were shifted. However, Kleinschmidt and Jaeger also observed a previously undocumented property of these adaptive changes: shifts in the exposure distribution had less than proportional (sublinear) effect on shifts in PSE (Figure 1C). While this finding—recently replicated in one more experiment (Dave Kleinschmidt, 2020, Experiment 4)—is broadly compatible with the hypothesis of distributional learning, it points to important not well-understood constraints on
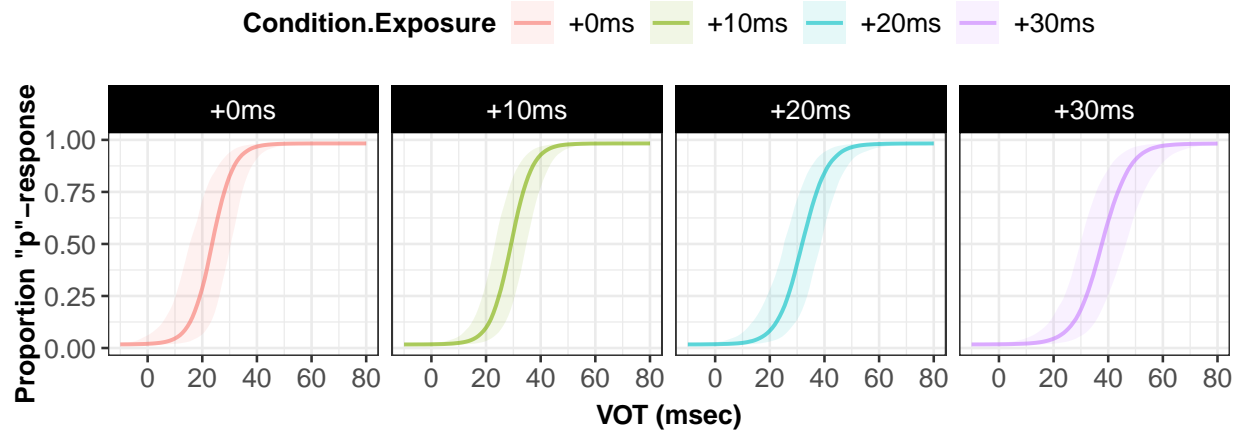
54 adaptive speech perception.

Figure 1. Design and results of D. F. Kleinschmidt and Jaeger (2016) replotted. **Panel A:** Different groups of participants were exposed to different shifts in the mean VOT of /b/ and /p/. **Panel B:** Categorisation functions of individual participants depending on the exposure condition (shift in VOT means of /b/ and /p/). For reference, the black dashed line shows the categorisation function of the 0-shift condition. The colored dashed lines shows the categorisation function expected for an ideal observer that has fully learned the exposure distributions. **Panel C:** Mean and 95% CI of participants' points of subjective equality (PSEs), relative to the PSE of the ideal observers.

55   For example, influential *models* of adaptive speech perception predict proportional, rather

56 than sublinear, shifts (for proof, see SI **??**). This is the case both for incremental Bayesian

57 belief-updating model (D. F. Kleinschmidt & Jaeger, 2011) and general purpose normalization

58 accounts (McMurray & Jongman, 2011)—models that have been found to explain listeners'

59 behavior well in experiments with less substantial changes in exposure. There are, however,

60 proposals that can accommodate this finding. Some proposals distinguish between two types of

61 mechanisms that might underlie representational changes, *model learning* and *model selection*

62 (**xie2018?**). The former refers to the learning of a new category representations—for example,

63 learning a new generative model for the talker (D. F. Kleinschmidt & Jaeger, 2015, pt. II) or

64 storage of new talker-specific exemplars (Sumner, 2011; **johnson1997?**). Xie and colleagues

65 hypothesized that this process might be much slower than is often assumed in the literature,

66 potentially requiring multiple days of exposure and memory consolidation during sleep (see also

67 **fenn2013?**; **tamminen2012?**; **xie2018sleep?**). Rapid adaptation that occurs within minutes of

68 exposure might instead be achieved by selecting between *existing* talker-specific representations

69 that were learned from previous speech input—e.g., previously learned talker-specific generative

70 models (see mixture model in D. F. Kleinschmidt & Jaeger, 2015, pp. 180–181) or previously

71 stored exemplars from other talkers (**johnson1997?**). Model learning and model selection both

72 offer explanations for the sublinear effects observed in D. F. Kleinschmidt and Jaeger (2016). But

73 they suggest different predictions for the evolution of this effect over the course of exposure.

74   Under the hypothesis of model learning, sublinear shifts in PSEs can be explained by

75 assuming a hierarchical prior over talker-specific generative models ($p(\Theta)$ in D. F. Kleinschmidt &

76 Jaeger, 2015, p. 180). This prior would 'shrink' adaptation towards listeners' priors—similar to

77 the effect of random by-subject or by-item effects in generalized linear mixed-effect models, which

78 shrink group-level effect estimates towards the population mean of the data (**bates?**). Critically,

79 as long as these priors attribute non-zero probability to even extreme shifts (e.g., the type of

80 Gaussian prior used in mixed-effects models), this predicts listeners' PSEs will continue to change

81 with increasing exposure until they have converged against the PSE that is ideal for the exposure

82 statistics. In contrast, the hypothesis of model selection predicts that rapid adaptation is more

83 strongly constrained by previous experience: listeners can only adapt their categorisation

84 functions up to a point that corresponds to (a mixture of) previously experienced talker-specific

85 generative models. Figure 2 visualizes the contrasting predictions of model learning and selection

86 for incremental adaptation.

87   To test these predictions, we revise the standard paradigm used to investigate distributional

88 learning in speech perception. Previous work has employed 'batch testing' designs, in which

89 changes in categorisation responses are assessed only after extended exposure to hundreds of trials

90 or by averaging over extended exposure (e.g., Clayards et al., 2008; Idemaru & Holt, 2011; D. F.

91 Kleinschmidt & Jaeger, 2016; Munson, 2011; Nixon et al., 2016; Theodore & Monto, 2019;

92 **harmon2018?**; **idemaru2021?**). These designs are well-suited to investigate cumulative effects

93 of exposure but are less so to identify constraints on rapidly unfolding incremental adaptation. To

94 be able to detect both incremental and cumulative effects of exposure, within and across exposure

95 conditions, we employed the repeated exposure-test design shown in Figure 3.

96   A secondary aim of the present study was to ameliorate possible concerns about the

97 ecological validity of research on distributional learning. The pioneering works that inspired the

98 present study employed highly unnatural sounding stimuli that were clearly identifiable as robotic
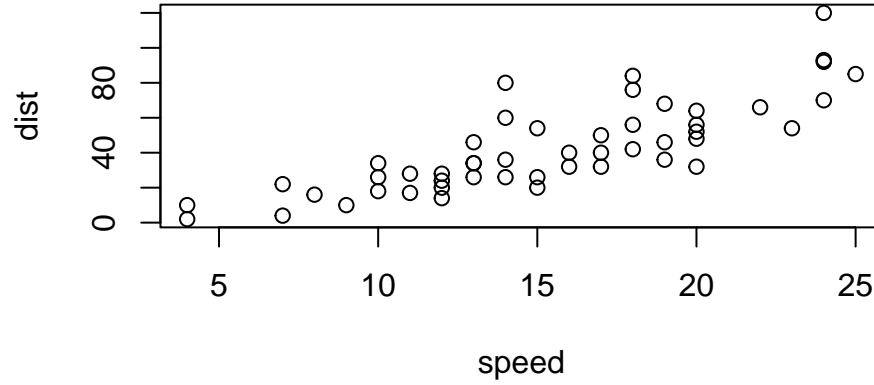
Figure 2. Contrasting predictions of model learning and model selection hypotheses about the incremental effects of exposure on listeners' categorisation function. Both hypothesis predict incremental adaptation towards the statistics of the input, as well as constraints on this adaptation. The two hypotheses differ, however, in that model selection predicts a hard limit on how far listeners' can adapt during initial encounters with an unfamiliar talker.
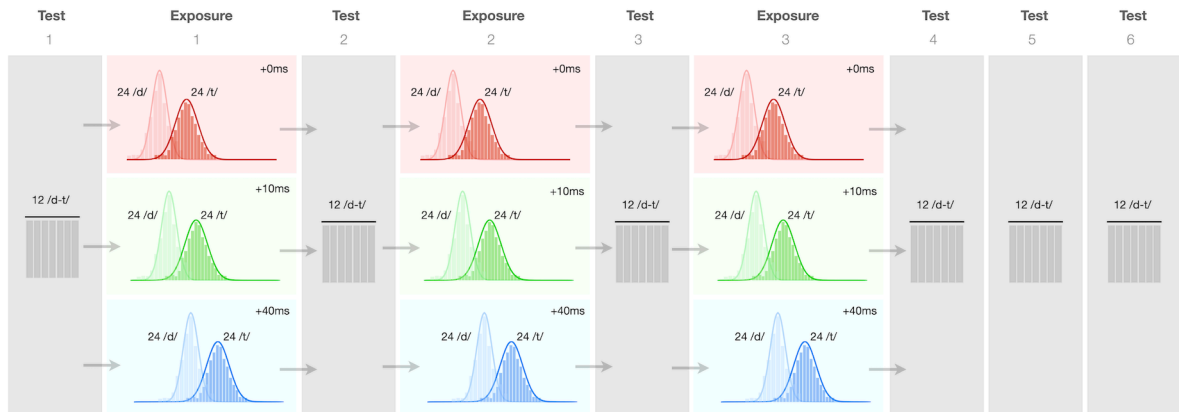
Figure 3. Exposure-test design of the experiment. Test blocks presented identical stimuli within and across conditions

speech (Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016). These studies also followed the

majority of research on distributional learning in language (e.g., **maye2003?**; **pajak2012?**) and

designed rather than *sampled* the exposure distributions. As a consequence, exposure

distributions in these experiments tend to be symmetrically balanced around the category

means—unlike in everyday speech input. Indeed, all of the works we follow here further used

categories with *identical* variances (e.g., identical variance along VOT for /b/ and /p/, Clayards

et al., 2008; D. F. Kleinschmidt & Jaeger, 2016; or /g/ and /k/, Theodore & Monto, 2019). This,

too, is highly atypical for everyday speech input (**chodroff2017structure?**;

lisker-abrahamson1964?). The present study takes several modest steps to address these

issues, with the goal to improve the ecological validity of our stimuli and exposure distributions.

All data and code for this article can be downloaded from XXX. The article is written in R

markdown, allowing readers to replicate our analyses with the press of a button using freely

available software (R, R Core Team, 2021; RStudio Team, 2020), while changing any of the

parameters of our models (see SI, **??**).

2 Experiment

The use of test blocks that repeat the same stimuli across blocks and exposure conditions deviates

from previous work (Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016; Theodore &

Monto, 2019). This design feature allowed us to assess how increasing exposure affects listeners'

perception without making strong assumptions about the nature of these changes (e.g., linear

changes across trials). We kept test blocks short for two reasons. First, previous work has found

that repeated testing over uniform test continua can reduce or undo the effects of informative

exposure (Liu & Jaeger, 2018, 2019; **cummings202X?**). Second, since we held test stimuli

constant across exposure conditions, the distribution—and thus the relative unexpectedness—of

test stimuli differed to different degrees from the three exposure distributions. By keeping tests

short relative exposure (12 vs. 48 trials), we aimed to minimize the influence of test trials on

adaptation. The final three test blocks were intended to ameliorate the potential risks of this novel

design: in case adaptation remains stable despite repeated testing, those additional test blocks

126 were meant to provide additional statistical power to detect the effects of cumulative exposure.

2.1 Methods

2.1.1 Participants

129 We recruited 126 participants from the Prolific crowdsourcing platform. We used Prolific's

130 pre-screening to limit the experiment to participants (1) of US nationality, (2) who reported to be

131 English speaking monolinguals, and (3) had not previously participated in any experiment from

132 our lab on Prolific. Prior to the start of the experiment, participants had to confirm that they (4)

133 had spent the first 10 years of their life in the US, (5) were in a quiet place and free from

134 distractions, and (6) wore in-ear or over-the-ears headphones that cost at least $15. An additional

135 115 participants loaded the experiment but did not start or complete it.[1]

136   Participants took an average of 31.6 minutes to complete the experiment (SD = 20

137 minutes) and were remunerated $8.00/hour. An optional post-experiment survey recorded

138 participant demographics using NIH prescribed categories, including participant sex (59 = female,

139 60 = male, 3 = NA), age (mean = NA years; 95% quantiles = 20-62.1 years), race (6 = Black, 31

140 = White, 85 = NA), and ethnicity (6 = Hispanic, 113 = Non-Hispanic, 3 = NA).

141   Participants' responses were collected via Javascript developed by the Human Language

142 Processing Lab at the University of Rochester (**JSEXP?**) and stored via Proliferate developed at,

143 and hosted by, the ALPs lab at Stanford University (**schuster?**).

2.1.2 Materials

145 We recorded 8 tokens each of four minimal word pairs (*dill/till*, *dim/tim*, *din/tin*, and *dip/tip*)

146 from a 23-year-old, female L1-US English talker from New Hampshire, judged to have a "general

147 American" accent. In addition to these critical minimal pairs we also recorded three words that

148 did not did not contain any stop consonant sounds ("flare", "share", and "rare"). These word

149 recordings were used for catch trials. Stimulus intensity was normalized to 70 dB sound pressure

---

[1] Unlike in lab-based experiments, for which participants' right to stop the experiment at any point is costly (both in terms of physical effort and perceived social cost), exercising this right in web-based experiments is essentially cost free—in particular, if exercised early in the experiment.

150 level for all recordings.

151   The critical minimal pair recordings were used to create four VOT continua using a script

152 (Winn, 2020) in Praat (**praat?**). This approach resulted in continuum steps that sound natural

153 (unlike the highly robotic-sounding stimuli employed in Clayards et al., 2008; D. F. Kleinschmidt

154 & Jaeger, 2016). A post-experiment survey asked participants: "*Did you notice anything in*

155 *particular about how the speaker pronounced the different words (e.g. till, dill, etc.)?*" No

156 participant reported that the stimuli sounded unnatural. The procedure also maintained the

157 natural correlations between the most important cues to word-initial stop-voicing in L1-US

158 English (VOT, F0, and vowel duration). Specifically, the F0 at vowel onset of each stimulus was

159 set to respect the linear relation with VOT observed in the original recordings of the talker. The

160 duration of the vowel was set to follow the natural trade-off relation with VOT (Allen & Miller,

161 1999). Further details on the recording and resynthesis procedure are provided in the

162 supplementary information (SI, **??**).

163   The VOTs generated for each continuum ranged from -100 to +130 msec in 5 msec steps.[2]

164 A norming experiment (N = 24 participants) reported in the SI (**??**) was used to select the three

165 minimal pair continua that elicited the most similar categorization responses (*dill-till*, *din-tin*, and

166 *dip-tip*). These three continua were used to create the exposure conditions shown in Figure 3.

2.1.3 Procedure

168 At the start of the experiment, participants acknowledged that they met all requirements and

169 provided consent, as per the Research Subjects Review Board of the University of Rochester.

170 Participants also had to pass a headphone test (**REF?**), and were instructed to not change the

171 volume throughout the experiment. Following instructions, participants completed 234

172 two-alternative forced-choice categorisation trials (Figure 4). Participants were instructed that

173 they would hear a female talker say a single word on each trial, and were asked to select which

174 word they heard. Participants were asked to listen carefully and answer as quickly and as

---

[2] We follow previous work (Dave Kleinschmidt, 2020; **OTHERS?**) and refer to prevoicing as negative VOTs though we note that prevoicing is perhaps better conceived of as a separate phonetic feature (for discussion, see **REF?**). In L1-US English, prevoicing occurs on about 20%-48% of word-initial voiced stops and ~0% of voiceless stops (**lisker-abramson1967?**; **smith1978?**).

175 accurately as possible. They were also alerted to the fact that the recordings were subtly different
176 and therefore may sound repetitive.
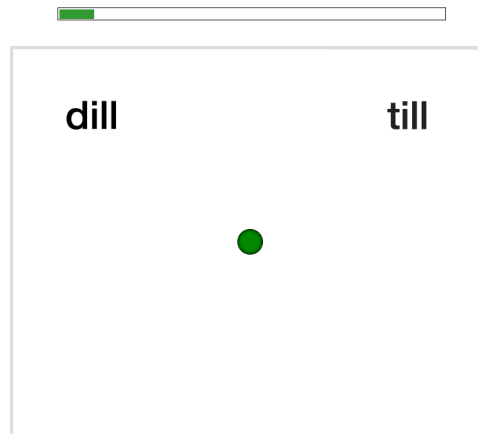
Figure 4. Example trial display. When the green button turned bright green, participants had to click on it to play the recording.

177   Unbeknownst to participants, the 234 trials were split into exposure (54 trials each) and
178 test blocks (12 trials each). Participants were given the opportunity to take breaks after every 60
179 trials, which was always during an exposure block. Finally, participants completed an exit survey
180 and an optional demographics survey.

181   *Test blocks.*   The experiment started with a test block. Test blocks were identical within
182 and across conditions, always including 12 minimal pair trials assessing participants'
183 categorization at 12 different VOTs (-5, 5, 15, 25, 30, 35, 40, 45, 50, 55, 65, 70 msec). A uniform
184 distribution over VOTs was chosen to maximize the statistical power to determine participants'
185 categorisation function. The assignment of VOTs to minimal pair continua was randomized for
186 each participant, while counter-balancing it within and across test blocks. Each minimal pair
187 appear equally often within each test block (four times), and each minimal pair appear with each
188 VOT equally often (twice) across all six test blocks (and no more than once per test block).

189   Each trial started with a dark-shaded green fixation dot being displayed. At 500ms from
190 trial onset, two minimal pair words appeared on the screen, as shown in Figure **??**. At 1000ms
191 from trial onset, the fixation dot would turn bright green and participants had to click on the dot
192 to play the recording. This was meant to reduce trial-to-trial correlations by resetting the mouse
193 pointer to the center of the screen at the start of each trial. Participants responded by clicking on

194 the word they heard and the next trial would begin.

195 *Exposure blocks.* Each exposure block consisted of 24 /d/ and 24 /t/ trials, as well as 6

196 catch trials that served as a check on participant attention throughout the experiment (2

197 instances for each of three combinations of the three catch recordings). With a total of 144 trials,

198 exposure was substantially shorter than in similar previous experiments (cf. 228 trials in Clayards

199 et al., 2008; 222 trials in Dave Kleinschmidt, 2020; 2 x 236 trials, Theodore & Monto, 2019; 456

200 trials, Nixon et al., 2016).

201 The distribution of VOTs across the 48 /d/-/t/ trials depended on the exposure condition.

202 Specifically, we first created a *baseline* condition. Although not critical to the purpose of the

203 experiment, we aimed for the VOT distribution in this condition to closely resemble participants'

204 prior expectations for a 'typical' female talker of L1-US English (for details, see SI, **??**). The

205 mean and standard deviations for /d/ along VOT were set 5 msecs and 50 msecs, respectively.

206 The mean and standard deviations for /t/ were set 80 msecs and 270 msecs, respectively. To

207 create more realistic VOT distributions, we *sampled* from the intended VOT distribution (top row

208 of Figure 5). This creates distributions that more closely resemble the type of distributional input

209 listeners experience in everyday speech perception, deviating from previous work, which exposed

210 listeners to highly unnatural fully symmetric samples (Clayards et al., 2008; Dave Kleinschmidt,

211 2020; D. F. Kleinschmidt & Jaeger, 2016).

212 Half of the /d/ and half of the /t/ trials were labeled, the other half was unlabeled

213 (paralleling one of the conditions in D. Kleinschmidt, Raizada, & Jaeger, 2015). Unlabeled trials

214 were identical to test trials except that the distribution of VOTs across those trials was bimodal

215 (rather than uniform), and determined by the exposure condition.[3] Labeled trials instead

216 presented two response options with identical stop onsets (e.g., *din* and *dill*). This effectively

217 labeled the input as belonging to the intended category (e.g., /d/).

218 Next, we created the two additional exposure conditions by shifting these VOT

---

[3] Previous studies have estimated changes in participants' categorisation responses by analyzing responses on unlabeled exposure trials (e.g., Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016; Theodore & Monto, 2019). This approach compares responses across different values of acoustic-phonetic cues (since the exposure inputs differed by exposure condition), so that assumptions baked into the analysis approach (e.g., linearity along the acoustic-phonetic continuum) can potentially bias the results. Here we avoid this issue by holding test stimuli constant (see also Dave Kleinschmidt, 2020, Experiment 4).

distributions by +10 or +40 msecs (see Figure 5). This approach exposes participants to

heterogenous approximations of normally distributed VOTs for /d/ and /t/ that varied across

blocks, while holding all aspects of the input constant across conditions except for the shift in

VOT. The order of trials was randomized within each block and participant, with the constraint

that no more than two catch trials would occur in a row. Participants were randomly assigned to

one of 3 (exposure condition) x 3 (block order) x 2 (placement of response options) lists.
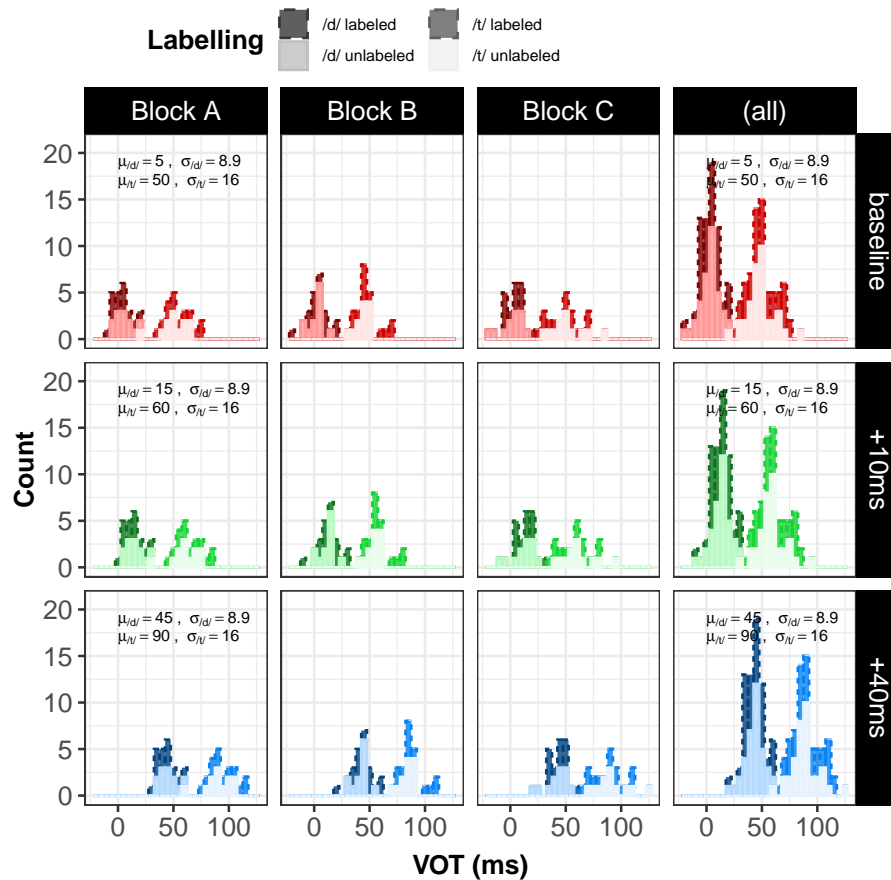

Figure 5. Histogram of VOTs across the 48 trials of all three exposure blocks by exposure condition. The order of blocks was counter-balanced across participants.

2.1.4   Exclusions

Due to data transfer errors 4 participants' data were not stored and therefore excluded from

analysis. We further excluded from analysis participants who committed more than 3 errors out

of the 18 catch trials ($<83\%$ accuracy, N = 1), participants who committed more than 4 errors

out of the 72 labelled trials (<94% accuracy, N = 0), participants with an average reaction time more than three standard deviations from the mean of the by-participant means (N =), participants who had atypical categorisation functions at the start of the experiment (N = 2, see SI, **??** for details), and participants who reported not to have used headphones (N = 0). This left for analysis 17,136 exposure and 8,568 test observations from 119 participants (94% of total), evenly split across the three exposure conditions.

2.2   Results

Figure 6 summarizes the results. Panels A and B show participants' categorisation responses during exposure and test blocks, along with the categorisation function estimated from those responses via a mixed-effects psychometric model that we describe next. These panels facilitate comparison between exposure conditions within each block. Panels C and D show the slope and point of subject equality (PSE)—i.e., the point at which participants are equally likely to respond "d" and "t"—of the categorisation function across blocks and conditions. These panels facilitate comparison across blocks within each exposure condition.

2.2.1   Analysis approach

We analyzed participants' categorisation responses during exposure and test blocks in two separate Bayesian mixed-effects psychometric models, fit using brms (Bürkner, 2017) in R (R Core Team, 2021; RStudio Team, 2020, for details, see SI, **??**). These models account for attentional lapses while estimating participants' categorisation functions. Failing to account for attentional lapses—while commonplace in research on speech perception (but see Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (Prins, 2012; Wichmann & Hill, 2001). For the present experiment, however, lapse rates were negligible (0.9%, 95%-CI: 0.4 to 1.5%), and all results replicate in simple mixed-effects logistic regressions (Jaeger, 2008).

2.2.2 Does exposure affect participants' categorisations (averaging across all blocks)?

Here we focus on the test blocks, which were identical within and across exposure conditions. Analyses of the exposure blocks are reported in the SI (**??**), and replicate all effects found in the test blocks. We first assessed whether the exposure conditions had the expected effects across all test blocks *relative to each other*. Unsurprisingly, participants were more likely to respond "t" the larger the VOT ($\hat{\beta} = 15.68$, $90\%-\text{CI} = [13.149, 18.4]$, $BF = 7999$, $p_{posterior} = 1$). Critically, exposure affects participants' categorisation responses in the expected direction. Marginalizing across all blocks, participants in the +40 condition were less likely to respond "t" than participants in the +10 condition ($\hat{\beta} = -2.43$, $90\%-\text{CI} = [-3.541, -1.363]$, $BF = 443.4$, $p_{posterior} = 0.998$) or the baseline condition ($\hat{\beta} = -3.39$, $90\%-\text{CI} = [-4.969, -1.93]$, $BF = 332.3$, $p_{posterior} = 0.997$). There was also evidence—albeit less decisive—that participants in the +10 condition were less likely to respond "t" than participants in the baseline condition ($\hat{\beta} = -0.97$, $90\%-\text{CI} = [-2.241, 0.298]$, $BF = 9.2$, $p_{posterior} = 0.902$). That is, the +10 and +40 conditions resulted in categorisation functions that were shifted rightwards compared to the baseline condition, as also visible in Figures 6.

This replicates previous findings that exposure to changed VOT distributions changes listeners' categorization responses (for /b/-/p/: Clayards et al., 2008; Dave Kleinschmidt, 2020; D. F. Kleinschmidt & Jaeger, 2016; for /g/-/k/, Theodore & Monto, 2019). Having established that exposure affected categorization, we turn to the questions of primary interest.

1. How quickly does exposure begin to affect participants' responses? And do they dissipate with repeated testing?
2. Are the shifts in categorisation behaviour proportional to the differences between the exposure conditions?
3. [model learning vs. selection]

[MORE HERE]
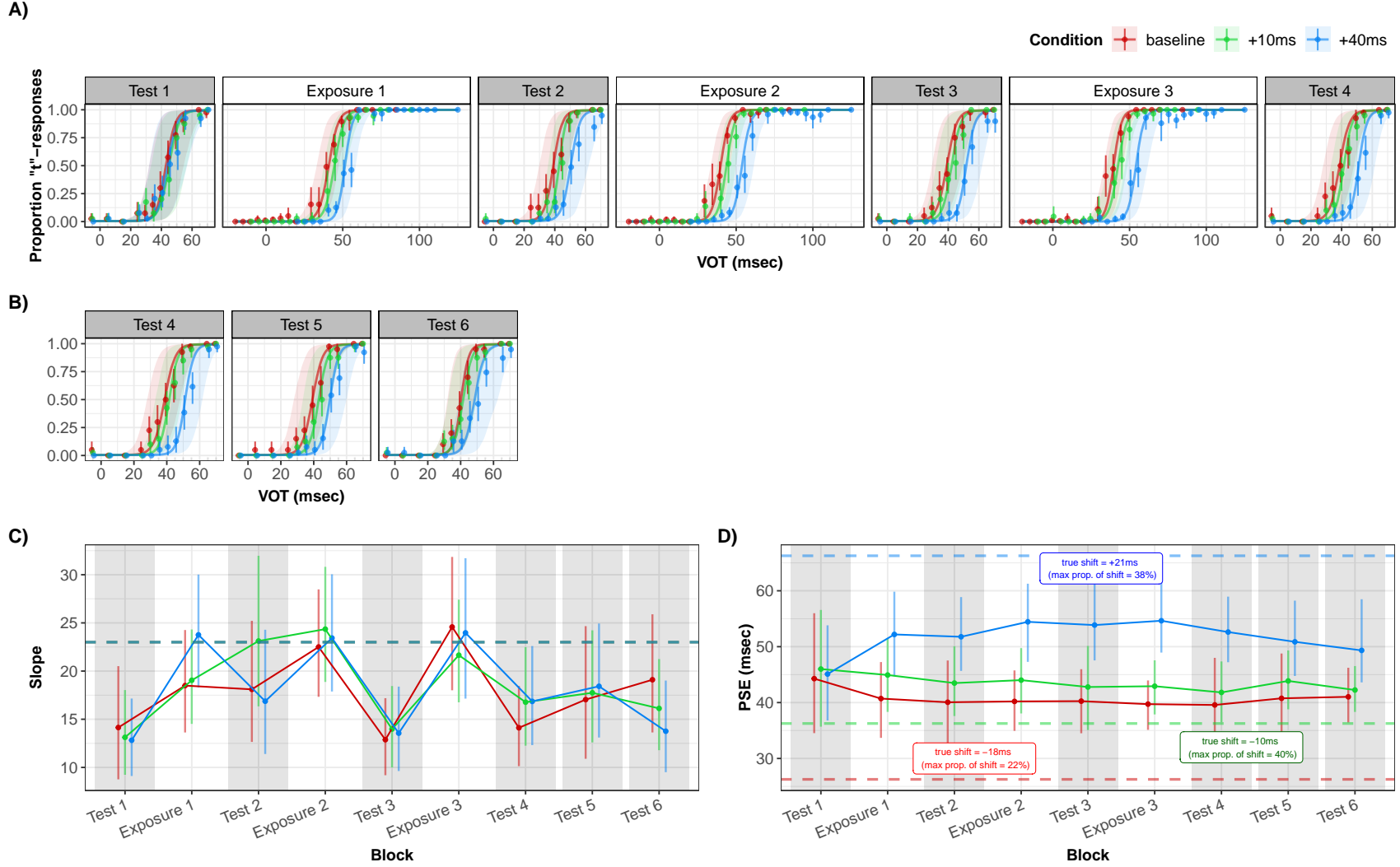
Figure 6. Summary of results. **Panel A:** Changes in listeners psychometric categorisation functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks (only unlabelled trials were included in the analysis of exposure blocks since labelled trials provide no information about listeners' categorization function). Point ranges indicate the mean proportion of "t"-responses and their 95% bootstrapped CI. Lines and shaded intervals show the MAP predictions and 95% posterior CIs of a Bayesian mixed-effects psychometric model fit to participants' responses. **Panel B:** Same as Panel A but for the final three test blocks without intervening exposure. Test 4 is shown as part of both Panels A and B. **Panels C & D:** Changes across blocks in the slope and boundary (point-of-subjective-equality, PSE) of the categorisation functions shown in Panels A-B. Point ranges represent the posterior means and their 95% CI. Dashed reference lines show the intercepts and PSEs that naive (non-rational) learner would be expected to converge against after sufficient exposure (an ideal observer model that knows the exposure distributions).

280   Next, we compared the effects of exposure conditions for all test blocks with intervening

281 exposure. These effects can be compared from three complementing perspectives. First, we can

282 compare how exposure affects listeners' categorisation responses relative to other exposure

283 conditions. Second, we can compare how exposure affects listeners' categorisation responses

284 within each condition relative to listeners' responses prior to any exposure. Third, we can

285 compare changes in listeners' responses to those expected from an ideal observer that has fully

286 learned the exposure distributions.

2.2.3 Comparing across exposure conditions: How quickly does exposure begin to affect participants' responses?

289 Figure 6A suggests that differences between exposure conditions emerged early in the experiment:

290 already in Test 2, listener's categorisation functions seem to be shifted rightwards (larger PSEs)

291 in the +40 condition compared to the +10 condition, and in the +10 condition compared to the

292 baseline condition. This is confirmed by the Bayesian hypothesis tests summarized in Table 1.

293 Prior to any exposure, during Test 1, participants' responses did not differ across exposure

294 condition. After exposure to only 24 /d/ and 24 /t/ stimuli, during Test 2, participants'

295 responses differed between exposure conditions. The difference between the +40 condition and

296 the +10 or baseline condition kept increasing with exposure up to Test 4. Additional hypothesis

297 tests in Table 2 show that the change from Test 1 to 2 was largest (BF = 27.8), followed by the

298 change from Test 2 to 3 (BF = 19.2), with only minimal changes from Test 3 to 4 (BF = 1.7).

299 Qualitatively paralleling the changes across blocks for the +40 condition, the change in the

300 difference between the +10 and baseline conditions was largest from Test 1 to 2 (BF = 13.5), and

301 then somewhat decreased from Test 2 to Test 4 (BFs < 4).

302   Tables 1 and 2 also summarize the consequences of repeated testing. The difference in the

303 PSE decreased from Test 4 to 6, both for the +40 compared to the +10 condition

304 ($\hat{\beta} = 1.98, \ 90\%-\text{CI} = [-0.418, 4.338], \ BF = 12.2, \ p_{posterior} = 0.924$) and the +10 compared to

305 the baseline condition ($\hat{\beta} = 0.93, \ 90\%-\text{CI} = [-0.921, 2.908], \ BF = 4.3, \ p_{posterior} = 0.811$, see

306 also Figure 6B & D). On the final test block, the +10 condition did not differ any longer from the

307 baseline condition. Only the differences between the +40 condition relative to the +10 and

308 baseline conditions persisted, albeit substantially reduced compared to Test 4.

309   This replicates previous findings that repeated testing over uniform test continua can undo

310 the effects of exposure (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019), and extends

311 them from perceptual recalibration paradigms to distributional learning paradigms (see also Dave

312 Kleinschmidt, 2020). One important methodological consequence of these findings is that longer

313 test phases do not necessarily increase the statistical power to detect effects of adaptation (unless

314 analyses take the effects of repeated testing into account, as in the approach developed in Liu &

315 Jaeger, 2018). Analyses that average across all test tokens—as remains the norm—are bound to

316 systematically underestimate the adaptivity of human speech perception.

2.2.4   Comparing within exposure conditions: How quickly does exposure begin to affect participants' responses?

319 Second, we can compare how exposure affects listeners' categorisation reponses within each

320 condition relative to listeners' responses prior to any exposure. These changes are summarised for

321 the slope and PSE in Figure 6C & D, respectively. This visualization makes apparent an aspect of

322 participants' behavior that were not readily apparent in the statistical comparisons we have

323 summarized so far: while the PSEs for the +40 and +10 conditions were shifted rightwards

324 compared to the baseline condition, both the +10 and the baseline condition actually shift

325 leftwards relative to their pre-exposure starting point in Test 1. To understand this pattern, it is

326 helpful to relate our exposure conditions to the distribution of VOT in listeners' prior experience.

327 Figure 7 shows the mean and covariance of our exposure conditions relative to the distribution of

328 VOT by talkers of L1-US English (based on Chodroff & Wilson, 2017). This comparison offers an

329 explanation as to why the baseline condition (and to some extent the +10 condition) shift

330 leftwards with increasing exposure, whereas the +40 condition shifts rightwards: relative to

331 listeners' prior experience our baseline condition actually presented lower-than-expected category

332 means; of our three exposure conditions, only the +40 condition presented larger-than-expected

333 category means. That is, once we take into account how our exposure conditions relate to

334 listeners' prior experience, both the direction of changes from Test 1 to 4 *within* each exposure

335 condition, and the direction of differences *between* exposure conditions receive an explanation.

336 **2.2.5   Constraints on cumulative changes**

337 Most relevant to the purpose of the present study, Figure 6D also suggests that there are strong

338 constraints on listeners' ability to adapt. To some extent, this already follows from the analysis of

339 how the differences between exposure conditions change with exposure.

340   XXX

341   Second, the reason for the slight decrease in the difference between the +10 and baseline

342 conditions observed in Tables **??** and **??** (visible in Figure 6D as the decreasing difference between

343 the green and red line) is *not* due to a reversal of the effects in the +10 condition. Rather, both

344 conditions are changing in the same direction but the baseline condition stops changing after Test

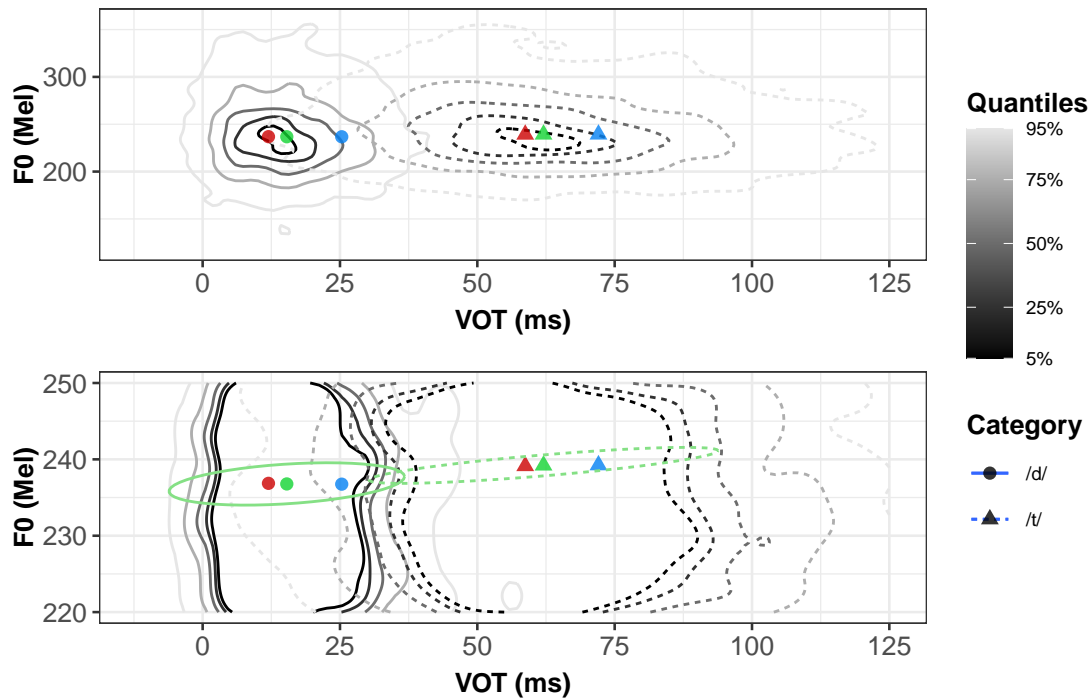345 2, which brings the +10 condition increasingly closer to the baseline condition.

Figure 7. Placement of exposure stimuli relative to an estimate of typical phonetic distributions for XXX word-initial /d/ and /t/ productions in L1-US English (based on 92 talkers in Chodroff & Wilson, 2017). The outermost contour of each category shows the 95% density quantile. Points show the category means of the exposure condition, the green ellipse shows the covariance of the +10 exposure condition (covariance was identical across conditions).

Table 1

*When did exposure begin to affect participants' categorization responses? When, if ever, were
these changes undone with repeated testing? This table summarizes the simple effects of the
exposure conditions for each test block.*

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Test block 1 (pre-exposure)					
+10 vs. baseline	-0.39	0.94	[-2.096, 1.403]	1.99	0.66
+40 vs. +10	0.20	0.86	[-1.359, 1.849]	0.68	0.40
+40 vs. baseline	-0.19	1.11	[-2.377, 2.041]	1.32	0.57
Test block 2					
+10 vs. baseline	-2.12	1.12	[-4.334, -0.109]	22.12	0.96
+40 vs. +10	-2.10	1.21	[-4.333, 0.071]	17.35	0.95
+40 vs. baseline	-4.22	1.47	[-7.048, -1.624]	80.63	0.99
Test block 3					
+10 vs. baseline	-0.88	0.69	[-2.244, 0.417]	7.98	0.89
+40 vs. +10	-3.26	0.96	[-5.164, -1.624]	169.21	0.99
+40 vs. baseline	-4.15	1.11	[-6.371, -2.226]	162.26	0.99
Test block 4					
+10 vs. baseline	-1.08	0.99	[-3.017, 0.947]	5.46	0.84
+40 vs. +10	-4.02	1.09	[-6.043, -2.284]	420.05	1.00
+40 vs. baseline	-5.10	1.43	[-7.839, -2.542]	132.33	0.99
Test block 5 (no additional exposure)					
+10 vs. baseline	-1.50	0.86	[-3.08, 0.086]	16.24	0.94
+40 vs. +10	-2.98	1.08	[-5.01, -1.205]	130.15	0.99
+40 vs. baseline	-4.10	1.52	[-6.811, -1.436]	73.77	0.99
Test block 6 (no additional exposure)					
+10 vs. baseline	-0.14	0.88	[-1.829, 1.456]	1.28	0.56
+40 vs. +10	-2.03	0.91	[-3.852, -0.396]	34.71	0.97
+40 vs. baseline	-3.15	1.39	[-5.754, -0.515]	31.39	0.97

3   General discussion

- discuss consequences of findings for other accounts (decision-making; normalization)

- discuss fact that test stimuli deviate from exposure stimuli to different extent. on the one
  hand, it's just 1/4 of all trials. on the other hand, we do see relatively systematic changes in
  slopes each time we test. so there is evidence that even these 12 trials can affect
  categorisation slopes (though it is worth keeping in mind that this is a comparison across
  different sets of stimuli). could this explain shrinkage? unlikely since it wasn't the case in
  kleinschmidt and jaeger. could it explain the constraint on adaptation? that's less clear. we

Table 2

Was there incremental change from test block 1 to 4? Did these changes dissipate with repeated testing from block 4 to 6? This table summarizes the interactions between exposure condition and block, whether the differences between exposure conditions changed from test block to test block.

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Difference in +10 vs. baseline					
Block 1 to 2: increased Δ_{PSE}	-1.40	0.92	[-3.065, 0.199]	13.52	0.93
Block 2 to 3: increased Δ_{PSE}	0.85	0.98	[-1.113, 2.775]	0.25	0.20
Block 3 to 4: increased Δ_{PSE}	-0.01	0.92	[-1.838, 1.885]	1.02	0.50
Block 1 to 4: increased Δ_{PSE}	-0.58	1.54	[-3.652, 2.483]	1.82	0.64
Block 4 to 5: decreased Δ_{PSE}	-0.38	0.71	[-1.734, 1.091]	0.42	0.30
Block 5 to 6: decreased Δ_{PSE}	1.25	0.77	[-0.143, 2.723]	13.95	0.93
Block 4 to 6: decreased Δ_{PSE}	0.86	0.97	[-0.921, 2.908]	4.30	0.81
Difference in +40 vs. +10					
Block 1 to 2: increased Δ_{PSE}	-2.05	1.03	[-3.89, -0.231]	27.78	0.96
Block 2 to 3: increased Δ_{PSE}	-1.79	1.06	[-3.688, -0.001]	19.15	0.95
Block 3 to 4: increased Δ_{PSE}	-0.39	1.18	[-2.629, 1.624]	1.70	0.63
Block 1 to 4: increased Δ_{PSE}	-4.28	1.53	[-7.158, -1.722]	101.56	0.99
Block 4 to 5: decreased Δ_{PSE}	1.41	1.07	[-0.541, 3.319]	8.66	0.90
Block 5 to 6: decreased Δ_{PSE}	0.60	0.94	[-1.271, 2.311]	2.79	0.74
Block 4 to 6: decreased Δ_{PSE}	1.99	1.25	[-0.418, 4.338]	12.24	0.92

354   can, however, compare the relative mean of exposure and test.

355   • could some form of moving window with historical decay explain the findings? On the one

356   hand if the moving window is very small, that would not explain why we do see some

357   *cumulative* changes across blocks (window must be at least $48 + 12 = 60$ trials). on the

358   other hand, the qualitative changes in the PSEs and slopes suggest that 12 trials can be

359   enough to change some aspects of the categorisation function. it's thus *possible* that

360   something that ways recent input much more strongly but also considers less recent input

361   beyond 48 trials might explain the overall pattern.

362   • discuss potential that observed adaptation maximizes accuracy under the choice rule. use

363   psychometric function fit during unlabeled exposure trials to calculate *accuracy* (not

364   likelihood) on labeled trials under criterion and under proportional matching decision rules.

365   compare against accuracy if ideal observers categorization functions are used instead.

Table 3

Effects of repeated testing (test blocks 4 to 6)

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Difference in +10 vs. baseline					
Predicted change from test block 4 to 5	-0.38	0.71	[-1.734, 1.091]	0.42	0.30
Predicted change from test block 5 to 6	1.25	0.77	[-0.143, 2.723]	13.95	0.93
Predicted total change from test block 4 to 6	0.86	0.97	[-0.921, 2.908]	4.30	0.81
Difference in +40 vs. +10					
Predicted change from test block 4 to 5	1.41	1.07	[-0.541, 3.319]	8.66	0.90
Predicted change from test block 5 to 6	0.60	0.94	[-1.271, 2.311]	2.79	0.74
Predicted total change from test block 4 to 6	1.99	1.25	[-0.418, 4.338]	12.24	0.92

3.1 Methodological advances that can move the field forward

4 References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4), 2031–2039.

Apfelbaum, K., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin and Review*, *22*, 916–943. https://doi.org/10.3758/s13423-014-0783-2

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE*, *6*, 1–12.

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

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

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization:

Table 4

Is the shift in +40 from baseline proportional to the magnitude of shift in the exposure distribution?

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Test block 2					
+40 vs. baseline < 4x +10 vs. baseline	4.24	4.0	[-2.563, 11.755]	6.00	0.86
Test block 3					
+40 vs. baseline < 4x +10 vs. baseline	-0.66	2.6	[-5.316, 4.068]	0.66	0.40
Test block 4					
+40 vs. baseline < 4x +10 vs. baseline	-0.84	3.4	[-7.448, 5.528]	0.67	0.40

Covariation of stop consonant VOT in american english. *Journal of Phonetics*, *61*, 30–47.

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

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. https://doi.org/https://doi.org/10.1016/j.cognition.2008.04.004

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

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology. Human Perception and Performance*, *37*, 1939–1956. https://doi.org/10.1037/a0025641

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kleinschmidt, Dave. (2020). *What constrains distributional learning in adults?*

Kleinschmidt, D. F., & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. *ACL Workshop on Cognitive Modeling and Computational Linguistics.*

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. https://doi.org/https://doi.org/10.1037/a0038695

Kleinschmidt, D. F., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? *CogSci*.

Kleinschmidt, D., Raizada, R., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci15)*. Austin, TX: Cognitive Science Society.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70.

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *45*, 1562–1588. https://doi.org/10.1037/xhp0000693

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–246. https://doi.org/10.1037/a0022325.What

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing* (University of Iowa). University of Iowa. Retrieved from http://ir.uiowa.edu/etd/2747

Nixon, J. S., Rij, J. van, Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125. https://doi.org/10.1016/j.jml.2016.03.005

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, *12*(6), 25–25.

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

RStudio Team. (2020). *RStudio: Integrated development environment for r*. Boston, MA:

RStudio, PBC. Retrieved from http://www.rstudio.com/

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

Tan, M., Xie, X., & Jaeger, T. F. (2021). Using rational models to understand experiments on accent adaptation. *Frontiers in Psychology*, *12*, 1–19. https://doi.org/10.3389/fpsyg.2021.676271

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review*, *26*(3), 985–992. https://doi.org/https://doi.org/10.3758/s13423-018-1551-5

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852–866.

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*.

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