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

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

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

[2] Brain and Cognitive Sciences, University of Rochester

[3] Computer Science, University of Rochester

Author Note

We are grateful to ### ommitted for review ###

Correspondence concerning this article should be addressed to Maryann Tan, Department of Bilingualism, Stockholm University, Sweden. E-mail: maryann.tan@biling.su.se

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

1 TO-DO

1.1 Highest priority

- MARYANN
- KJ16 plot: fix legends, fix y-axis to conform to current convention, add 0 reference line *typical talker + IO references lines (predicted PSEs)
- main plot:
  - remove vertical gridlines
  - reference dashed lines: put labels for e.g. "+21.x ms (100%)" on the right edge.
  - change point estimate to a label with "+/-XX.X%" (rounded to 1 digit). data frame must have all info necessary to calculate data for that label geom.
- fit nested model: Condition / (block*VOT). Sample prior = "yes". This is to make the argument of block-to-block change within each condition.
  - make a hypothesis table that summarises the main effect of block for each exposure condition
- Non-parametric density plot.
  - Apply correction for vowel duration!
- Try an add line to table 2 to separate the unlearning hypothesis from the others (low priority). Add +40 vs baseline sub-heading

1.2 Medium priority

- MARYANN
- add a dashed gray reference line to Figure 6C/D, showing the slope and PSE based on io fitted to Chodroff & Wilson?
- KJ16 plot: include point ranges in PSE comparison plot

- Fix a lot of the outstanding XXXes. Fill in the references – in library.bib
- Heterogenous normal distribution
- fix appearance of annotations in histogram plot
- FLORIAN
- think about table 1 and 2: how to change the wording on tables to actually refer to intercepts rather than PSEs or change the figures? Changing current representations of analyses to improve intuitive-ity.
- write overview of results
- restructure results presentation.
- write SI sections with proofs

1.2.1 Lower Priority

- MARYANN
- standardize “ms” vs. “msec”
- Decide on PSE vs. category boundary
- standardize BE vs. AE spelling (categoriz/sation, label(l)ed, synthesiz/sed etc.)
- Figure 2: A small figure to anticipate the type of format in the main figure. Exposure is x-axis (less to more exposure). Under one prediciton you keep growing 00 prediction o BBU model. The other one is plataeuing and converging against something.
- Florian
- compare IBBU predictions over blocks with human behavioural data

1.3 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; Davis & Sohoglu, 2020; Idemaru & Holt, 2011; D. Kleinschmidt & Jaeger, 2015; for review, Schertz & Clare, 2020; **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, Beck, Lu, & Pouget, 2011; Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2012; Nixon, Rij, Mok, Baayen, & Chen, 2016; Theodore & Monto, 2019; **idemaru2021?**; **munson2011-thesis?**; **tan2021?**; 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 (D. Kleinschmidt, 2020, Experiment 4)—

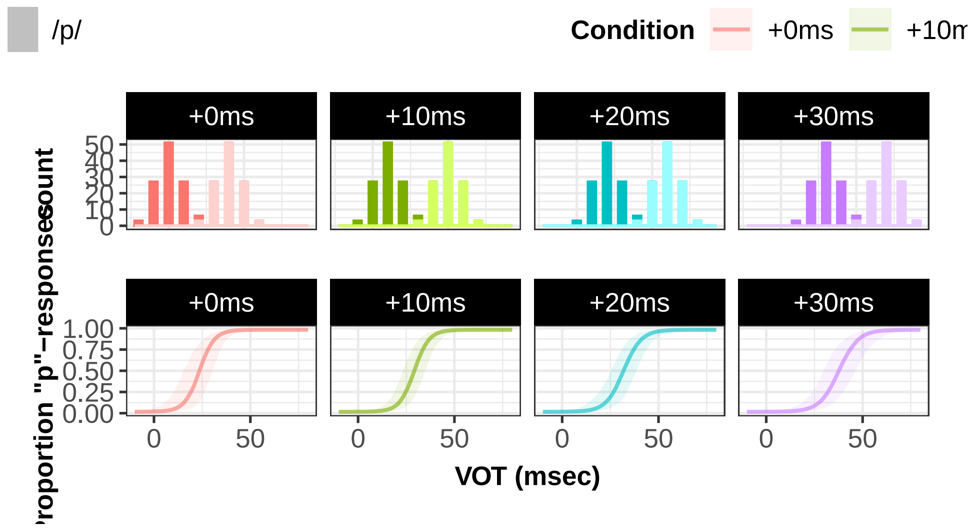

Figure 1. (ref:kleinschmidt-jaeger-2016-replotted)

For example, influential *models* of adaptive speech perception predict proportional, rather than sublinear, shifts (for proof, see SI **??**). This is the case both for incremental Bayesian belief-updating model (**kleinschmidt-jaeger2011?**) and general purpose normalization accounts (McMurray & Jongman, 2011)—models that have been found to explain listeners' behavior well in experiments with less substantial changes in exposure. There are, however, proposals that can accommodate this finding. Some proposals distinguish between two types of mechanisms that might underlie representational changes, *model learning* and *model selection* (Xie et al., 2018, p. 229). The former refers to the learning of a new category representations—for example, learning a new generative model for the talker (D. Kleinschmidt & Jaeger, 2015, pt. II) or storage of new talker-specific exemplars (**johnson1997?**; **sumner2011?**). Xie and colleagues hypothesized that this process might be much slower than is often assumed in the literature, potentially requiring multiple days of exposure and memory consolidation during sleep (Tamminen, Davis, Merkx, & Rastle, 2012; Xie, Earle, & Myers, 2018; see also **fenn2013?**). Rapid adaptation that occurs within minutes of exposure might instead be achieved by selecting between *existing* talker-specific representations that were learned from previous speech input—e.g., previously learned talker-specific generative models (see mixture model in D. Kleinschmidt & Jaeger, 2015, pp. 180–181) or previously stored exemplars from other talkers (**johnson1997?**). Model learning and model selection both offer explanations for the sublinear effects observed in D. F. Kleinschmidt and Jaeger (2016). But they suggest different predictions for the evolution of this effect over the course of exposure.

Under the hypothesis of model learning, sublinear shifts in PSEs can be explained by assuming a hierarchical prior over talker-specific generative models ($p(\Theta)$ in D. Kleinschmidt & Jaeger, 2015, p. 180). This prior would 'shrink' adaptation towards listeners' priors—similar to the effect of random by-subject or by-item effects in generalized linear mixed-effect models, which shrink group-level effect estimates towards the population mean of the data (Baayen, Davidson, & Bates, 2008). Critically, as long as these priors attribute non-zero probability to even extreme shifts (e.g., the type of Gaussian prior used in mixed-effects models), this predicts listeners' PSEs will continue to change with increasing exposure until they have converged against the PSE that is ideal for the exposure statistics. In contrast, the hypothesis of model selection predicts that rapid adaptation is more strongly constrained by previous experience: listeners can only adapt their categorisation functions up to a point that corresponds to (a mixture of) previously experienced talker-specific generative models. Figure 2 visualizes the contrasting predictions of model learning and selection for incremental adaptation.

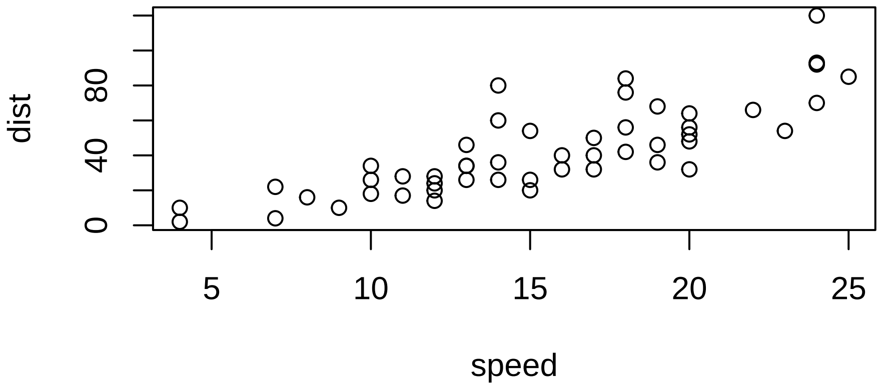

Figure 2. Contrasting predictions of model learning and model selection hypotheses about the incremental effects of exposure on listeners' categorisation function. Both hypothesis predict incremental adaptation towards the statistics of the input, as well as constraints on this adaptation. The two hypotheses differ, however, in that model selection predicts a hard limit on how far listeners' can adapt during initial encounters with an unfamiliar talker.

To test these predictions, we revise the standard paradigm used to investigate distributional learning in speech perception. Previous work has employed 'batch testing' designs, in which changes in categorisation responses are assessed only after extended exposure to hundreds of trials or by averaging over extended exposure (e.g., Clayards et al., 2008; Idemaru & Holt, 2011; D. F. Kleinschmidt & Jaeger, 2016; Nixon et al., 2016; Theodore & Monto, 2019; **harmon2018?**; **idemaru2021?**; **munson2011?**). These designs are well-suited to investigate cumulative effects of exposure but are less so to identify constraints on rapidly unfolding incremental adaptation. To be able to detect both incremental and cumulative effects of exposure, within and across exposure conditions, we employed the repeated exposure-test design shown in Figure 3.

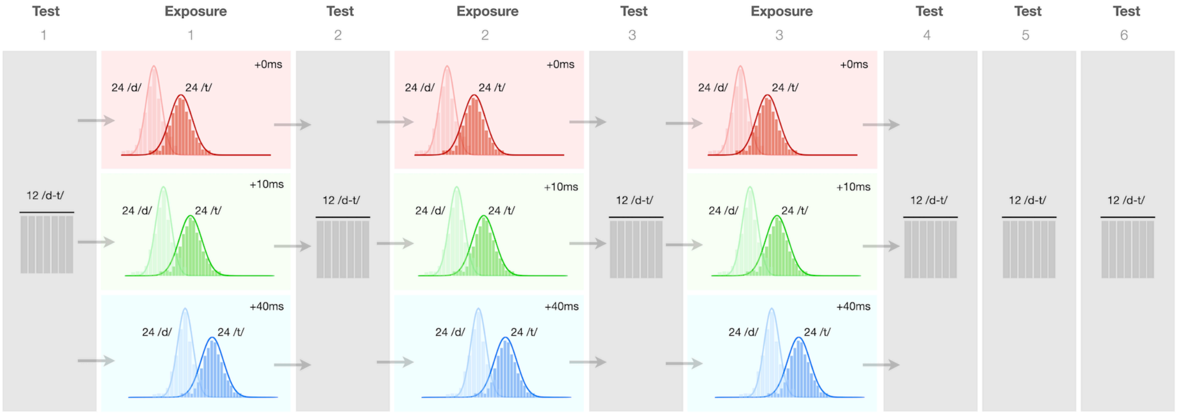

Figure 3. Exposure-test design of the experiment. Test blocks presented identical stimuli within and across conditions

A secondary aim of the present study was to ameliorate possible concerns about the ecological validity of research on distributional learning. The pioneering works that inspired the present study employed highly unnatural sounding stimuli that were clearly identifiable as robotic 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., Maye, Werker, & Gerken, 2002; Pajak & Levy, 2012) 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 (Chodroff & Wilson, 2018;

lisker-abrahamson1964?). The present study takes several modest steps to address these issues, to assess whether the observations made by D. F. Kleinschmidt and Jaeger (2016) replicate for ecologically more valid 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?**; **RStudio?**), 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-jaeger2018?**; **liu-jaeger2019?**; **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 were meant to provide additional statistical power to detect the effects of cumulative exposure.

2.1 Methods

2.1.1 Participants

We recruited 126 participants from the Prolific crowdsourcing platform. We used Prolific's pre-screening to limit the experiment to participants (1) of US nationality, (2) who reported to be

English speaking monolinguals, and (3) had not previously participated in any experiment from our lab on Prolific. Prior to the start of the experiment, participants had to confirm that they (4) had spent the first 10 years of their life in the US, (5) were in a quiet place and free from distractions, and (6) wore in-ear or over-the-ears headphones that cost at least $15. An additional 115 participants loaded the experiment but did not start or complete it.[1]

Participants took an average of 31.6 minutes to complete the experiment (SD = 20 minutes) and were remunerated $8.00/hour. An optional post-experiment survey recorded participant demographics using NIH prescribed categories, including participant sex (59 = female, 60 = male, 3 = NA), age (mean = NA years; 95% quantiles = 20-62.1 years), race (6 = Black, 31 = White, 85 = NA), and ethnicity (6 = Hispanic, 113 = Non-Hispanic, 3 = NA).

Participants' responses were collected via Javascript developed by the Human Language Processing Lab at the University of Rochester (**JSEXP?**) and stored via Proliferate developed at, and hosted by, the ALPs lab at Stanford University (**schuster?**).

2.1.2 Materials

We recorded 8 tokens each of four minimal word pairs (*dill/till*, *dim/tim*, *din/tin*, and *dip/tip*) from a 23-year-old, female L1-US English talker from New Hampshire, judged to have a "general American" accent. In addition to these critical minimal pairs we also recorded three words that did not did not contain any stop consonant sounds ("flare", "share", and "rare"). These word recordings were used for catch trials. Stimulus intensity was normalized to 70 dB sound pressure level for all recordings.

The critical minimal pair recordings were used to create four VOT continua using a script (Winn, 2020) in Praat (**praat?**). This approach resulted in continuum steps that sound natural (unlike the highly robotic-sounding stimuli employed in Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016). A post-experiment survey asked participants: "*Did you notice anything in particular about how the speaker pronounced the different words (e.g. till, dill, etc.)?*" No

[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.

participant reported that the stimuli sounded unnatural. The procedure also maintained the natural correlations between the most important cues to word-initial stop-voicing in L1-US English (VOT, F0, and vowel duration). Specifically, the F0 at vowel onset of each stimulus was set to respect the linear relation with VOT observed in the original recordings of the talker. The duration of the vowel was set to follow the natural trade-off relation with VOT (Allen & Miller, 1999). Further details on the recording and resynthesis procedure are provided in the supplementary information (SI, **??**).

The VOTs generated for each continuum ranged from -100 to +130 msec in 5 msec steps.[2] A norming experiment (N = 24 participants) reported in the SI (**??**) was used to select the three minimal pair continua that elicited the most similar categorization responses (*dill-till*, *din-tin*, and *dip-tip*). These three continua were used to create the exposure conditions shown in Figure 3.

2.1.3 Procedure

At the start of the experiment, participants acknowledged that they met all requirements and provided consent, as per the Research Subjects Review Board of the University of Rochester. Participants also had to pass a headphone test (Woods, Siegel, Traer, & McDermott, 2017), and were instructed to not change the volume throughout the experiment. Following instructions, participants completed 234 two-alternative forced-choice categorisation trials (Figure 4). Participants were instructed that they would hear a female talker say a single word on each trial, and were asked to select which word they heard. Participants were asked to listen carefully and answer as quickly and as accurately as possible. They were also alerted to the fact that the recordings were subtly different and therefore may sound repetitive.

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

[2] We follow previous work (D. 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?**).

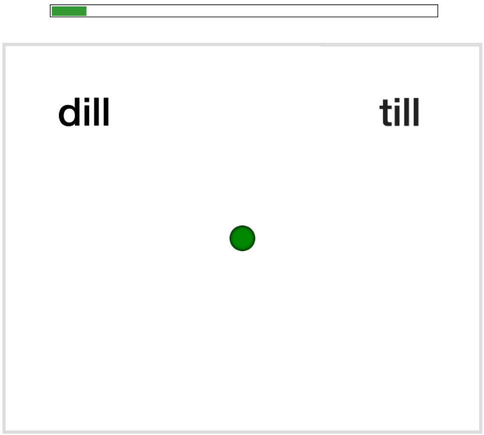

Figure 4. Example trial display. When the green button turned bright green, participants had to click on it to play the recording.

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

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

Exposure blocks. Each exposure block consisted of 24 /d/ and 24 /t/ trials, as well as 6 catch trials that served as a check on participant attention throughout the experiment (2 instances for each of three combinations of the three catch recordings). With a total of 144 trials, exposure was substantially shorter than in similar previous experiments (cf. 228 trials in Clayards et al., 2008; 222 trials in D. Kleinschmidt, 2020; 2 x 236 trials, Theodore & Monto, 2019; 456

trials, Nixon et al., 2016).

The distribution of VOTs across the 48 /d/-/t/ trials depended on the exposure condition. Specifically, we first created a *baseline* condition. Although not critical to the purpose of the experiment, we aimed for the VOT distribution in this condition to closely resemble participants' prior expectations for a 'typical' female talker of L1-US English (for details, see SI, **??**). The mean and standard deviations for /d/ along VOT were set 5 msecs and 50 msecs, respectively. The mean and standard deviations for /t/ were set 80 msecs and 270 msecs, respectively. To create more realistic VOT distributions, we *sampled* from the intended VOT distribution (top row of Figure 5). This creates distributions that more closely resemble the type of distributional input listeners experience in everyday speech perception, deviating from previous work, which exposed listeners to highly unnatural fully symmetric samples (Clayards et al., 2008; D. Kleinschmidt, 2020; D. F. Kleinschmidt & Jaeger, 2016).

Half of the /d/ and half of the /t/ trials were labeled, the other half was unlabeled (paralleling one of the conditions in D. F. Kleinschmidt, Raizada, & Jaeger, 2015). Unlabeled trials were identical to test trials except that the distribution of VOTs across those trials was bimodal (rather than uniform), and determined by the exposure condition.[3] Labeled trials instead presented two response options with identical stop onsets (e.g., *din* and *dill*). This effectively labeled the input as belonging to the intended category (e.g., /d/).

Next, we created the two additional exposure conditions by shifting these VOT distributions by +10 or +40 msecs (see Figure 5). This approach exposes participants to heterogeneous 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.

[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 D. Kleinschmidt, 2020, Experiment 4).

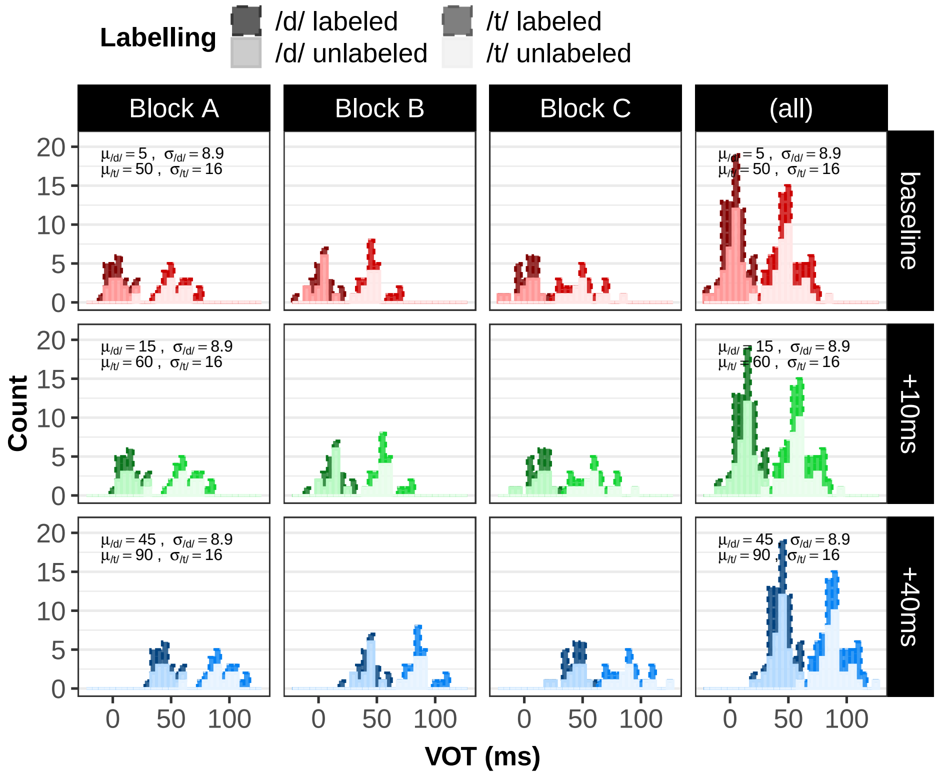

Figure 5. Histogram of voice onset times (VOTs) for each of the three exposure blocks A-C by trial type (/d/ or /t/, labeled or unlabeled) and exposure condition (baseline vs. +10 vs. +40). Each exposure block contained 12 labeled /d/, 12 labeled /t/, 12 unlabeled /d/, and 12 unlabeled /t/ trials, as well as 6 catch trials (not shown). Except for the shift in VOTs (+0, 10 or 40 msecs VOT to each trial), the VOT distribution of these trials was identical across exposure conditions. The order of exposure blocks A-C was counter-balanced across participants using a Latin-square design.

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

We analyzed participants' categorisation responses during exposure and test blocks in two separate Bayesian mixed-effects psychometric models, using brms (Bürkner, 2017) in R (**R?**; **RStudio?**, for details, see SI, **??**). Psychometric 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, 2011; 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).

Each psychometric model regressed participants' categorisation responses against the full factorial interaction of VOT, exposure condition, and block, while including the maximal random effect structure (see SI, **??**. Figure 6 summarizes the results that we describe in more detail next. Panels A and B show participants' categorisation responses during exposure and test blocks, along with the categorisation function estimated from those responses via the mixed-effects psychometric models. 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. 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 begin by presenting the overall effects, averaging across all test blocks. This part of our analysis matches previous work, which has focused on the overall effect of exposure across the entire experiment ('batch tests,' e.g., Clayards et al., 2008; Nixon et al., 2016; Theodore & Monto, 2019; **kleinschmidt2016?**) and/or during a single post-exposure test block (e.g., D. Kleinschmidt, 2020). Then we turn to the goals of this study—to characterize the incremental changes in participants' categorisation responses as a function of exposure and, in particular, to test 1) whether we replicate the sublinear effects of exposure observed in previous work under the ecologically more valid stimuli and distributions employed in the present work, and 2) whether we can begin to distinguish between the predictions of the model learning and selection hypotheses.

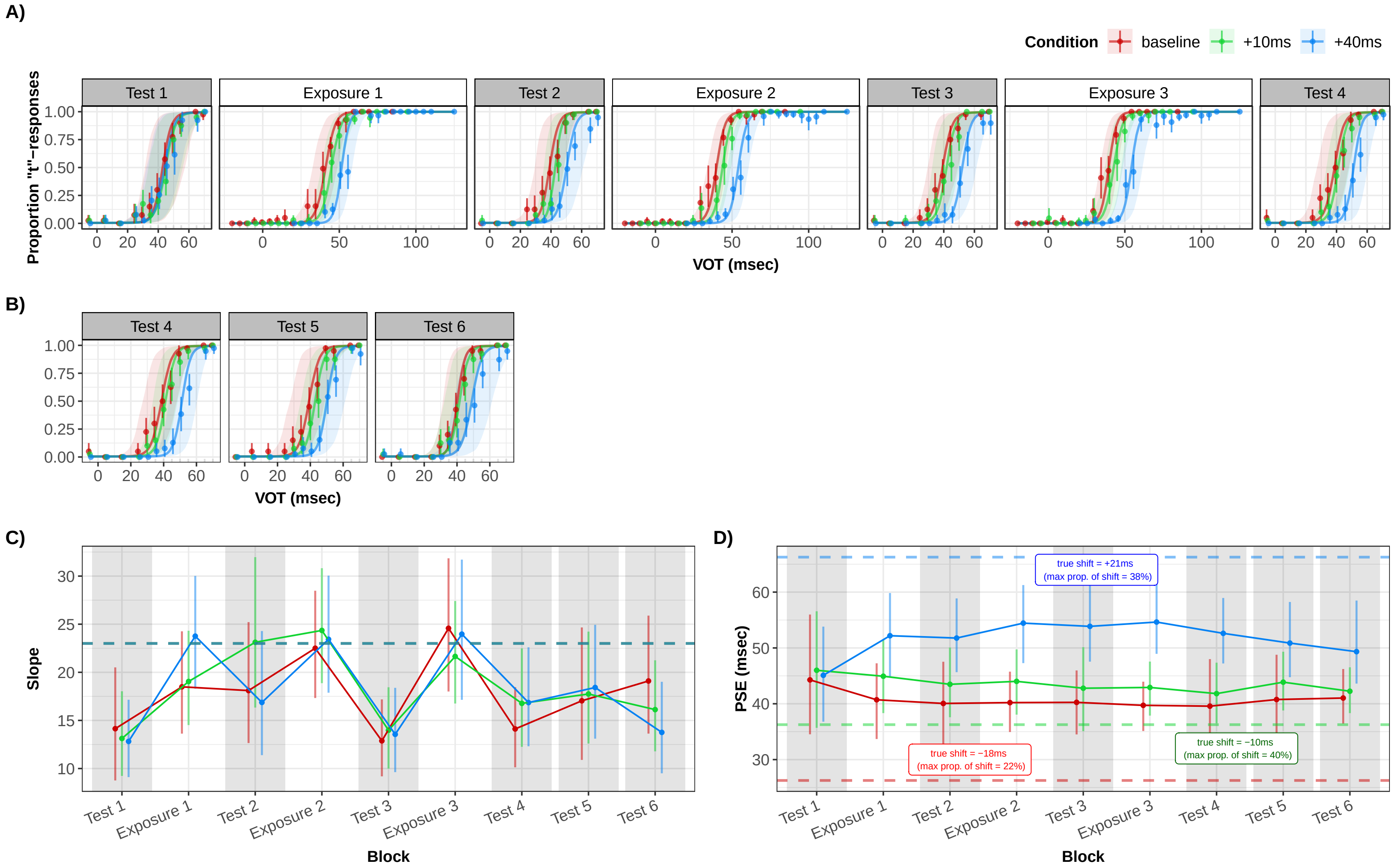

Figure 6. Summary of results. **Panel A:** Changes in listeners psychometric categorisation functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks (only unlabelled trials were included in the analysis of exposure blocks since labelled trials provide no information 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).

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

We first used the psychometric mixed-effects model to assess 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 ($\hat{\beta} = 15.68,\ 90\%{-}\mathrm{CI} = [13.149, 18.4],\ BF = 7999,\ p_{posterior} = 1$). 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 ($\hat{\beta} = -2.43,\ 90\%{-}\mathrm{CI} = [-3.541, -1.363],\ BF = 443.4,\ p_{posterior} = 0.998$) or the baseline condition ($\hat{\beta} = -3.39,\ 90\%{-}\mathrm{CI} = [-4.969, -1.93],\ BF = 332.3,\ p_{posterior} = 0.997$). 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 ($\hat{\beta} = -0.97,\ 90\%{-}\mathrm{CI} = [-2.241, 0.298],\ BF = 9.2,\ p_{posterior} = 0.902$). 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; D. 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. Incremental changes in participants' categorisation responses can be assessed from three mutually complementing perspectives. First, we compare how exposure affects listeners' categorisation responses relative to other exposure conditions. This tests how early in the experiment differences between exposure conditions began to emerge. Second, we compare how exposure affects listeners' categorisation responses within each condition relative to listeners' responses prior to any exposure. This assesses how the exposure conditions relate to participants' prior expectations. Most importantly, however, it tests the subtly different predictions of the model learning and selection hypotheses—whether changes in listeners' categorisation responses are strongly constrained. Third and finally, we compare changes in listeners' responses to those expected from an ideal observer that has fully learned the exposure distributions. This tests whether the sublinear effects observed in D. F. Kleinschmidt and Jaeger (2016) replicate in our repeated exposure-test paradigm with the improvements the present study makes to ecological validity.

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

Figure 6A suggests that differences between exposure conditions emerged early in the experiment: already in Test 2, listener's categorisation functions seem to be shifted rightwards (larger PSEs) in the +40 condition compared to the +10 condition, and in the +10 condition compared to the baseline condition. This is confirmed by the Bayesian hypothesis tests summarized in Table 1. Prior to any exposure, during Test 1, participants' responses did not differ across exposure condition (all BFs > XXX). After exposure to only 24 /d/ and 24 /t/ stimuli, during Test 2, participants' responses differed between exposure conditions (BFs > 17.35). The difference between the +40 condition and the +10 or baseline condition kept increasing with exposure up to Test 4. Additional hypothesis tests in Table 2 show that the change from Test 1 to 2 was largest (BF = 27.8), followed by the change from Test 2 to 3 (BF = 19.2), with only minimal changes from Test 3 to 4 (BF = 1.7). Qualitatively paralleling the changes across blocks for the +40 condition, the change in the difference between the +10 and baseline conditions was largest from Test 1 to 2 (BF = 13.5), and then somewhat decreased from Test 2 to Test 4 (BFs < 4). The comparison across exposure conditions thus suggests that changes in listeners' categorisation responses emerged quickly—indeed, they were present already *during* the first exposure block (see SI, **??**)—but then leveled off. The comparison across exposure conditions also yields one result that is, at first blush, surprising: while the difference between the +10 and the baseline condition emerged already after the first exposure block, this difference *de*creased, rather than increased, with additional exposure from Test 2 to 3 (see second row of Table 2). We return to this effect below.

Tables 1 and 2 also reveal the consequences of repeated testing. The difference between exposure conditions decreased from Test 4 to 6 (BFs > 4.3; see also Figure 6B & D). On the final test block, the +10 condition did not differ any longer from the baseline condition. Only the differences between the +40 condition relative to the +10 and baseline conditions persisted, albeit substantially reduced compared to Test 4. This pattern of results replicates previous findings that repeated testing over uniform test continua can undo the effects of exposure (Cummings & Theodore, 2023; **liu-jaeger2018?**; **liu-jaeger2019?**), and extends them from perceptual recalibration paradigms to distributional learning paradigms (see also D. Kleinschmidt, 2020). One important methodological consequence of these findings is that longer test phases do not necessarily increase the statistical power to detect effects of adaptation (unless analyses take the effects of repeated testing into account, as in the approach developed in **liu-jaeger2018?**). Analyses that average across all test tokens—as remains the norm—are bound to systematically underestimate the adaptivity of human speech perception.

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

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
Difference in +40 vs. baseline					
Block 1 to 2: increased Δ_{PSE}	-3.44	1.13	[-5.612, -1.371]	87.89	0.99
Block 2 to 3: increased Δ_{PSE}	-0.96	1.33	[-3.488, 1.549]	3.32	0.77
Block 3 to 4: increased Δ_{PSE}	-0.42	1.45	[-3.303, 2.249]	1.56	0.61
Block 1 to 4: increased Δ_{PSE}	-4.85	2.16	[-9.019, -0.955]	36.04	0.97
Block 4 to 5: decreased Δ_{PSE}	1.03	1.22	[-1.254, 3.372]	3.92	0.80
Block 5 to 6: decreased Δ_{PSE}	1.86	1.12	[-0.329, 3.955]	13.18	0.93
Block 4 to 6: decreased Δ_{PSE}	2.88	1.56	[-0.178, 5.939]	16.13	0.94

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

Next, we compared how exposure affects listeners' categorisation reponses within each condition relative to listeners' responses prior to any exposure. These changes are summarised for the slope and PSE in Figure 6C & D, respectively. This visualization makes apparent two aspects of participants' behavior that were not readily apparent in the statistical comparisons we have summarized so far. First, while the PSEs for the +40 and +10 conditions were shifted rightwards compared to the baseline condition, both the +10 and the baseline condition actually shift leftwards relative to their pre-exposure starting point in Test 1. This is confirmed by Bayesian hypothesis tests summarized in Table **??**.

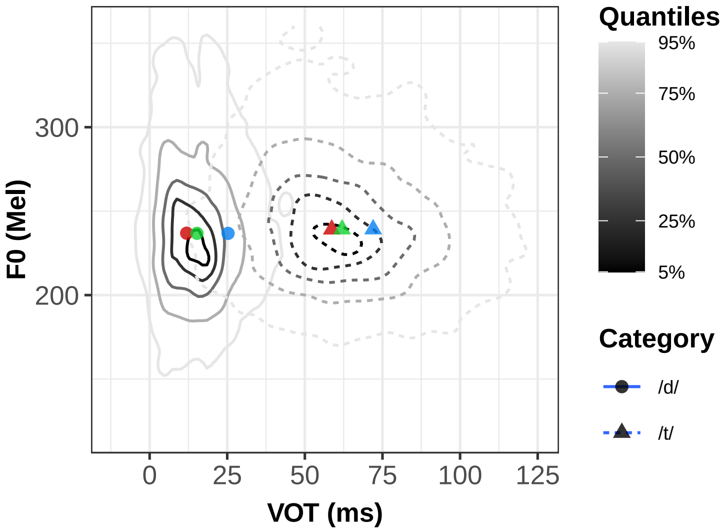

Figure 7. Placement of exposure stimuli relative to an estimate of typical phonetic distributions for XXX word-initial /d/ and /t/ productions in L1-US English (based on 92 talkers in Chodroff & Wilson, 2018). The outermost contour of each category shows the 95% density quantile. Points show the category means of the exposure condition.

To understand this pattern, it is helpful to relate our exposure conditions to the distribution of VOT in listeners' prior experience. Figure 7 shows the mean and covariance of our exposure conditions relative to the distribution of VOT by talkers of L1-US English (based on **chodroff-wilson2017?**). This comparison offers an explanation as to why the baseline condition (and to some extent the +10 condition) shift leftwards with increasing exposure, whereas the +40 condition shifts rightwards: relative to listeners' prior experience our baseline condition actually presented lower-than-expected category means; of our three exposure conditions, only the +40 condition presented larger-than-expected category means. That is, once we take into account how our exposure conditions relate to listeners' prior experience, both the direction of changes from Test 1 to 4 *within* each exposure condition, and the direction of differences *between* exposure conditions receive an explanation.

Second, the reason for the slight decrease in the difference between the +10 and baseline conditions observed in Tables **??** and 2 (visible in Figure 6D as the decreasing difference between the green and red line) is *not* due to a reversal of the effects in the +10 condition. Rather, both conditions are changing in the same direction but the baseline condition stops changing after Test 2, which reduces the difference between the +10 and baseline conditions (see Table **??**). The comparison across blocks thus suggests a rather uniform picture across all exposure conditions: participants' responses initially changed rapidly with exposure; with increasing exposure, these changes did not only slow down but seem to hit a hard constraint. Participants in the leftwards-shifted baseline condition did not exhibit any further changes in their categorisation responses beyond Test 2. Similarly, participants in the rightwards-shifted +40 condition did not exhibit any further changes in their categorisation responses beyond Test 3. Only participants in the leftward-shifted +10 condition still exhibit changes across blocks even form Test 3 to 4. But, perhaps tellingly, those participants also never reached the degree of shift that was evident in the baseline condition.

2.2.4 Constraints on cumulative changes

Finally, Figures 6C & D also compare participants' responses against those of an ideal observer that has fully learned the exposure distributions.

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 can, however, compare the relative mean of exposure and test.
- could some form of moving window with historical decay explain the findings? On the one hand if the moving window is very small, that would not explain why we do see some *cumulative* changes across blocks (window must be at least 48 + 12 = 60 trials). on the

other hand, the qualitative changes in the PSEs and slopes suggest that 12 trials can be enough to change some aspects of the categorisation function. it's thus *possible* that something that ways recent input much more strongly but also considers less recent input beyond 48 trials might explain the overall pattern.

- discuss potential that observed adaptation maximizes accuracy under the choice rule. use psychometric function fit during unlabeled exposure trials to calculate *accuracy* (not likelihood) on labeled trials under criterion and under proportional matching decision rules. compare against accuracy if ideal observers categorization functions are used instead.

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. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, *22*, 916–943.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bejjanki, V. R., Beck, J. M., Lu, Z.-L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, *14*(5), 642–648.

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. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2).

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.

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.

Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for bayesian inference in speech perception. *The Cognitive Neurosciences*, 177–189.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

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

Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*.

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

Kleinschmidt, D. F., Raizada, R. D., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. *CogSci*.

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

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*(2), 219.

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.

Pajak, B., & Levy, R. (2012). Distributional learning of L2 phonological categories by listeners with different language backgrounds. *Proceedings of the 36th Boston University Conference on Language Development*, *2*, 400–413. Cascadilla Press Somerville, MA.

Prins, N. (2011). The psychometric function: Why we should not, and need not, estimate the lapse rate. *Journal of Vision*, *11*(11), 1175–1175.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), e1521.

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112.

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review*, *26*, 985–992.

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.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064–2072.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210.

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22.

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