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

We investigate constraints on adaptive speech perception during the initial encounters with unfamiliar speech patterns. Such adaptive changes are now considered important to spoken language understanding, overcoming substantial cross-talker variability in the realization of speech categories. We present evidence from a novel incremental exposure-test paradigm to assess how previously experienced cross-talker variability guides (and thus constrains) listeners' adaptation. Specifically, we ask adaptation is constrained weakly—slower and sublinear, but continued adaptation with increasing exposure—or strongly—adaptation only up to a point, after which additional exposure has no benefits (at least not prior to, e.g., sleep). The results contribute to a proposed theoretical distinction between two hypotheses about the mechanisms underlying the intial moments of adaptation, model learning vs. model selection.

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
- REFIT THE EXPOSURE MODEL UNDER THE CORRECT DIFF CODING if it wasn't coded that way before
- edit Analysis Approach section in the SI
- Please read this carefully.
- TIME TO STOP MESSY CODING. Let's have a zero-tolerance policy for that from now on in the main working branch (i.e., you can do what you'd like in branches that aren't the main branch, but you canNOT merge without cleaning up first). It is a real time-sink for everyone else and makes it near impossible for me to effectively help.
  - on the main working branch, functions should be in functions.R, in a clearly named section (see existing examples).
- Input data file:
  - There shouldn't be multiple data files that you're loading. E.g., I don't understand why there is an exposure trials data file in addition to the main data file. It's just confusing. Let's not do things like that.
  - Have a script in your other repo (for your thesis) that does all the data importing, variable and value formatting, etc. The input data file experiment-results.csv should already contain all the information you (and others might need) and be in the format that you'd like it to be. That's the only data file that will be in your paper repo.
    - Think carefully about how to name variables consistently and create all variants of variables you might need in the paper, e.g., Response, Item.ExpectedResponse, Response.Category, Item.ExpectedResponse.Category, Response.Voiced, Item.ExpectedResponse.Voiced (etc. if you indeed need all of those; we definitely need the first two pairs of these).

    * Also if you have to consistently rename levels for plotting, please just changed them once in the script that creates the file. E.g., there's various places in which you deal with formatting the conditions and various names floating around (Shift0, 10, etc.; +0, +10, etc.; baseline, + 10 etc.). Pick one, do it at the top of the pipeline (i.e., in the input script). This will reduce the potential for error in your own coding, make your code in the main paper shorter, and it'll be much easier to read for others trying to follow your code (including me).
    * Remove all data formatting code from the paper Rmd. There should only be a single load line.
    * I've moved the code loading the chodroff data into the new pre-amble.R file. Consider doing the same for the experiment data. That way the data that we need throughout are available throughout.

- Clean up functions.R file:
  - PLEASE DO GET RID OF UNUSED FUNCTIONS. Search files for each function (cmd + shift + f). If it does not exist, remove it from functions.R
  - Use clearer function names. It often happens as a project develops that functions become ambiguous in their name. E.g., you have several functions that do similar things (like getting or plotting CIs from psychometric or IO models). Extend their names to be clear: e.g., compare get_CI to get_CI_from_ideal_observer; or make_CI to print_CI; or add_PSE_perception_median to add_PSE_median_to_plot (note how I also removed redundancy since PSEs are always about perception); etc. Rename the functions and use CMD + SHIFT + F to search and replace all mentions of those functions across all files.
  - Organize functions into sections with headings in functions.R

- Try to set local constants at top of chunk. e.g., Don't have stuff like empirical_means <- c(17, 62) in the middle of a chunk.

1.2 Medium priority

- MARYANN
- 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
- Combine data from exposure and test, use all together instead of coding block, code trial and code it as a smooth. That means using GAMM – that may require taking lapse (try it first without lapses because the GAMM takes care of the lapse. The RE will be expressed differently. It has to follow the GAMM syntax.) The primary thing we want to smooth over is "block", but could theoretically smooth over VOT and Block.
- 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

Adaptivity is a hallmark of human speech perception, supporting faster and more accurate speech recognition. When exposed to an unfamiliar accent, the processing difficulty listeners might initially experience tends to alleviate with exposure (Bradlow, Bassard, & Paller, 2023; e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Sidaras, Alexander, & Nygaard, 2009; Xie, Liu, & Jaeger, 2021; Xie et al., 2018). Research over the last few decades has made strides in identifying the conditions required for successful adaptation, its generalizability across talkers, and its longevity (for reviews, see Bent & Baese-Berk, 2021; Cummings & Theodore, 2023; **zheng-samuel2023?**). It is now clear that listeners' categorization function—the mapping from acoustic or phonetic inputs to linguistic categories and, ultimately, word meanings—changes based on the phonetic properties of recent input (e.g., Bertelson, Vroomen, & De Gelder, 2003; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Eisner & McQueen, 2005; Idemaru & Holt, 2011; Kraljic & Samuel, 2005; McMurray & Jongman, 2011; Norris, McQueen, & Cutler, 2003; Reinisch & Holt, 2014; **cole2011?**; **kurumada2013?**; **xie2018jep?**; for review, Schertz & Clare, 2020; Xie, Jaeger, & Kurumada, 2023). This has led to the development of stronger theories and models of adaptive speech perception that explicitly link the distribution of phonetic properties in recent speech input to changes in subsequent speech recognition (e.g., Apfelbaum & McMurray, 2015; Assmann & Nearey, 2007; Harmon, Idemaru, & Kapatsinski, 2019; Johnson, 1997; Kleinschmidt & Jaeger, 2015; Lancia & Winter, 2013; Magnuson et al., 2020; Sohoglu & Davis, 2016; Xie et al., 2023).

Previous work has typically framed questions as an 'either-or'—adaptation is either observed or not—consistent with the focus on identifying the necessary conditions for adaptation and generalization (see discussion in Cummings & Theodore, 2023). Recent reviews of the field instead emphasize the need to move towards stronger tests of existing theories, requiring the development of paradigms that support quantitative comparison to more strongly constrain the space of theoretical possibilities (Schertz & Clare, 2020; Xie et al., 2023; **baeseberk2018?**). This includes the need for data that characterize how adaptation develops *incrementally* as a function of exposure. While existing theories differ in important aspects, they share critical predictions about incremental adaptation that have remained largely untested: listeners' categorizations are predicted to change incrementally with exposure, and the direction and magnitude of that change should gradiently depend on (1) listeners' prior expectations based on previously experienced speech input from other talkers, and both (2a) the amount and (2b) distribution of phonetic evidence in the exposure input from the unfamiliar talker (for review, see Xie et al., 2023). We report initial results from a novel repeated exposure-test paradigm designed to test these predictions during the early moments of adaptation.

Figure 1 illustrates our approach. The experiment builds on computational and behavioral findings from separate lines of research on unsupervised distributional learning during speech perception (DL, Clayards et al., 2008; Kleinschmidt, 2020; Theodore & Monto, 2019), lexically- or visually-guided perceptual learning (LGPL, Cummings & Theodore, 2023; VGPL, Kleinschmidt & Jaeger, 2012; Vroomen, Linden, De Gelder, & Bertelson, 2007), and accent adaptation (AA, Hitczenko & Feldman, 2016; Tan, Xie, & Jaeger, 2021). These studies have complementing strengths that we seek to combine and extend. Following previous work on distributional learning in speech perception, we expose different groups of listeners to phonetic distributions that are shifted to different degrees (Bejjanki, Beck, Lu, & Pouget, 2011; Clayards et al., 2008; Kleinschmidt, Raizada, & Jaeger, 2015; Munson, 2011; Nixon, Rij, Mok, Baayen, & Chen, 2016; Theodore & Monto, 2019). Unlike this work, we incrementally assess changes in listeners' categorization from pre-exposure onward.

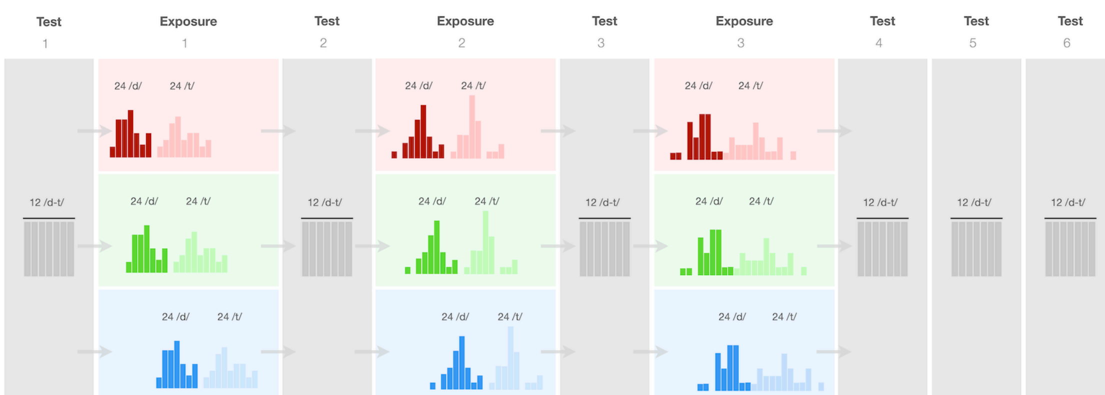

Figure 1. Exposure-test design of the experiment. Exposure conditions (rows) differed in the distribution of voice onset time (VOT), the primary phonetic cue to word-initial /d/ and /t/ in English (e.g., "dip" vs. "tip"). Test blocks assessed listeners' categorization functions over VOT stimuli that were held identical within and across conditions.

Following previous DL studies, we use phonetically manipulated stimuli. This gives researchers control over the distribution of acoustic-phonetic properties that listeners experience during exposure and test (unlike AA, LGPL, and VGPL paradigms). Such control is an important prerequisite for stronger tests of predictions (1) and (2a,b). For example, recent findings from LGPL and VGPL provide evidence in support of prediction (2a)—that the amount of phonetic evidence during exposure gradiently affects the magnitude of subsequent changes in listeners' categorization response (Cummings & Theodore, 2023; see also Liu & Jaeger, 2018, 2019). This includes some initial evidence that these changes accumulate incrementally (Kleinschmidt & Jaeger, 2012; Vroomen et al., 2007), in ways consistent with models of adaptive speech perception. LGPL and VGPL paradigms—at least as used traditionally—do, however, limit experimenters' control over the phonetic properties of the exposure stimuli: shifted sound instances are selected to be perceptually ambiguous (e.g., between "s" and "sh"), rather than to exhibit specific phonetic distributions. To the extent that LGPL and VGPL research has assessed the effects of phonetic properties on the degree of boundary shift following exposure, this has been limited to qualitative post-hoc analyses (Drouin, Theodore, & Myers, 2016; Kraljic & Samuel, 2007; Tzeng, Nygaard, & Theodore, 2021?). This makes it difficult to test predictions (1) and (2b) about the effects of phonetic distributions in prior and recent experience.

Support for prediction (2b) has thus primarily come from research in DL paradigms. In an important early study, Clayards et al. (2008) exposed two different groups of US English listeners to instances of "b" and "p" that differed in their distribution along the voice onset time continuum (VOT). VOT is the primary phonetic cue to word-initial /b/-/p/, /d/-/t/, /g/-/k/ in US English: the voiced category (e.g. /b/) is produced with lower VOT than the voiceless category (e.g., /p/). Clayards and colleagues held the VOT means of /b/ and /p/ constant between the two exposure groups, but manipulated whether both /b/ and /p/ had wide or narrow variance along VOT. Exposure was unlabeled: on any trial, listeners saw pictures of, e.g., bees and peas on the screen while hearing a synthesized recording along the "bees"-"peas" continuum (obtained by manipulating VOT). Listeners' task was to click on the picture corresponding to the word they heard. If listeners adapt by learning how /b/ and /p/ are distributed along VOT, listeners in the wide variance group were predicted to exhibit a more shallow categorization function than the narrow variance group. This is precisely what Clayards and colleagues found (see also Nixon et al., 2016; Theodore & Monto, 2019). Together with more recent findings from adaptation to natural accents (Hitczenko & Feldman, 2016; Tan et al., 2021; Xie, Buxó-Lugo, & Kurumada, 2021), this important finding suggests that the *outcome* of adaptation qualitatively follows the predictions of distributional learning models (e.g., exemplar theory, Johnson, 1997; ideal adaptors, Kleinschmidt & Jaeger, 2015). The findings in this line of work did, however, rely on tests that either averaged over, or followed, hundreds of trials of exposure. This leaves open how adaptation proceeds from the earliest moments of exposure—i.e., whether listeners' categorization behavior indeed changes in the way predicted by models of adaptive speech perception, developing from expectations based on previously experienced phonetic distributions to increasing integration of the phonetic distributions observed during exposure to the unfamiliar talker. It also leaves open whether potential constraints on the extent to which listeners' behavior changes with exposure (for initial evidence and discussion, see Cummings & Theodore, 2023; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016) reflect hard limits on adaptivity or simply reflect the incremental learning outcome—'how far the learner has gotten'—at the only point at which adaptation is assessed (i.e., following exposure).

The repeated exposure-test paradigm in Figure 1 begins to address these knowledge gaps. The experiment starts with a test block that assesses listeners' state prior to informative exposure—often assumed, but not tested, to be identical across exposure conditions. Additional intermittent tests—opaque to participants—then assess incremental changes up to the first 144 informative exposure trials. The use of physically identical test trials both across block within exposure conditions and across exposure conditions, we aim to facilitate assumption-free comparison of cumulative exposure effects (we additionally also measure adaptation during exposure). As we detail under Methods, the use of repeated testing deviates from previous work (Clayards et al., 2008; Harmon et al., 2019; Idemaru & Holt, 2011, 2020; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016; Munson, 2011; Nixon et al., 2016; Theodore & Monto, 2019), and is not without challenges. This design allows tests of prediction (2a) by comparing between participants, and of prediction (2b) by comparing within and across participants. The design also lets us assess how the joint effect exposure amount and exposure distributions—corresponding to predictions (2a) and (2b)—unfolds incrementally with exposure. And, by comparing the direction of adaptation not only across conditions, but also relative to the distribution of phonetic cues in listeners' prior experience, we can begin to assess prediction (1).

Finally, we took several modest steps towards addressing concerns about ecological validity that have been argued to limit the generalizability of DL results. This includes concerns about the ecological validity of both the stimuli and their distribution in the experiment (see discussion in **baseberk2018?**). For example, previous distributional learning studies have often used highly unnatural, 'robotic'-sounding, speech (but see Theodore & Monto, 2019). Beyond raising questions about what types of expectations listeners apply to such speech, these stimuli also failed to exhibit naturally occurring covariation between phonetic cues that listeners are known to expect (see, e.g., Idemaru & Holt, 2011; Schertz, Cho, Lotto, & Warner, 2016). We instead developed stimuli that both sound natural and exhibit the type of phonetic covariation that listeners expect from everyday speech perception. We return to these and additional steps we took to increase the ecological validity of the phonetic *distributions* under Methods.

All data and code for this article can be downloaded from https://osf.io/hxcy4/. Following Xie et al. (2023), both this article and its supplementary information (SI) are written in R markdown, allowing readers to replicate and validate our analyses with the press of a button using freely available software [R, R Core Team (2022); RStudio Team (2020); see also SI, **??**].

2 Methods

2.1 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' 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, S, 2020). Participants took an average of 31.6 minutes (SD = 20 minutes) to complete the experiment and were remunerated $8.00/hour. An optional post-experiment survey recorded participant demographics using NIH prescribed categories, including participant sex (female: 59, male: 60, declined to report: 3), age (mean = 38 years; SD = 12; 95% quantiles = 20-62.1 years), race (White: 31, Black: 6, declined to report: 85), and ethnicity (Non-Hispanic: 113, Hispanic: 6, declined to report: 3).

2.2 Materials

We recorded 8 tokens each of four minimal word pairs with word-initial /d/-/t/ (*dill/till*, *dim/tim*, *din/tin*, and *dip/tip*) from a 23-year-old, female L1-US English talker from New Hampshire. 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 ranging from -100 to +130 ms in 5 ms steps.[2] Continua were generated using a script (Winn, 2020) in Praat (Boersma & Weenink, 2022). This approach resulted in continuum steps that sound natural [unlike the highly robotic-sounding stimuli employed in previous work]. It 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,

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

[2] We follow previous work (Kleinschmidt, 2020; Lisker & Abramson, 1964) and refer to pre-voicing as negative VOTs though we note that pre-voicing is perhaps better conceived of as a separate phonetic feature (for discussion, see **REF?**). This distinction can, for example, be important when interpreting asymmetries in listeners' ability to adapt to left- vs. rightward shifts along the VOT continuum, an issue we return in the general discussion.

1999). Further details on the recording and resynthesis procedure are provided in the supplementary information (SI, **??**). 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 participant responded that the stimuli sounded unnatural. Perhaps more importanlty, analyses reported in the SI (**??**) found that participants exhibited few attentional lapses even in the first blocks of the experiment (< 1%). This is a marked improvement over previous studies with robotic sounding stimuli, which elicited high lapse rates at the start of the experiment (> 10%, Kleinschmidt, 2020). A norming experiment (N = 24 participants) reported in the SI (**??**) was used to select the three minimal pair continua that differed the least from each other in terms of the categorization responses they elicited (*dill-till*, *din-tin*, and *dip-tip*).

2.3 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 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 categorization trials. 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.

For the two-alternative forced-choice categorization trials, participants were instructed that they would hear a female talker say a single word on each trial, and had 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. 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 2. 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. Unbeknownst to participants, the 234 trials were split into three exposure blocks (54 trials each) and six test blocks (12 trials each, as shown in Figure 1).

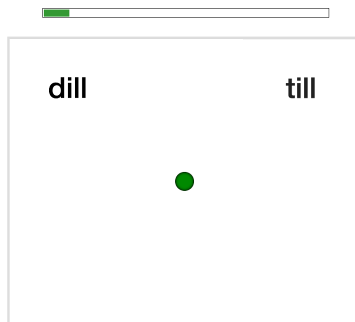

Figure 2. Example trial display. When the green button turned bright green, participants had to click on it to play the recording. The placement of response options was counter-balanced across participants.

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 ms). A uniform, rather than bimodal, distribution over VOTs was chosen to maximize the statistical power to determine participants' categorization function. Identical test blocks followed each exposure block to assess the effects of cumulative exposure. As alluded to in the introduction, the use of repeated testing introduces procedural challenges. These informed the decision to keep testing short. First, listeners' attention span is limited. Second, previous experiments within LGPL paradigms have found that repeated testing over uniform test continua can reduce or undo the effects of informative exposure (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019; Tzeng et al., 2021). Our design included two additional test blocks without intermittent exposure at the end of the experiment, in order to test whether repeated testing has similar effects in DL paradigms. Third, holding the distribution of test stimuli constant across exposure condition inevitably means that the relative unexpectedness of these test stimuli differs between the exposure conditions. Under some theories, this is expected to affect the information conveyed by test stimuli (Kleinschmidt & Jaeger, 2015; Sohoglu & Davis, 2016). By keeping tests short relative to exposure, we aimed to minimize the influence of test trials on adaptation while still being able to estimate changes in listeners categorization 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). The order of response options—whether the /d/-initial word appeared on the left or right of the screen (see Figure 2)—was held constant within each participant, and counter-balanced across participants.

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, and intermittent tests after 0, 48, and 96 critical trials, we assessed the effects of exposure at substantially earlier moments than in similar previous experiments (cf. 228 trials in Clayards et al., 2008; 222 trials in 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. 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 approximately resemble participants' prior expectations for a 'typical' female talker of L1-US English. Based on the norming experiment mentioned under *Materials*, we set the VOT means of 5ms for /d/ and 50ms for /t/ (for details, see SI, **??**). We took additional two steps to increase the ecological validity of the VOT distributions that deviate from similar previous work (Clayards et al., 2008; Idemaru & Holt, 2011, 2020; Kleinschmidt, 2020; Kleinschmidt et al., 2015). First, previous studies exposed each group of listeners to categories with identical variance. We instead set the variance for /d/ to 80 ms^2 VOT and for /t/ to 270 ms^2. This asymmetry reflects the natural distribution of VOT (**REF?**). Specifically, we set the variance of /d/ and /t/ based on a phonetic database of L1 US English word-initial /d/-/t/ productions (Chodroff & Wilson, 2017). Second, rather than to expose listeners to fully symmetric *designed* distributions that would never be experienced in everyday speech, we *randomly sampled* from the intended VOT distribution. The sampling-based approach instead creates VOT distributions that more closely resemble the type of speech input listeners experience outside of the lab (see top row of Figure 3). Specifically, we sampled VOTs for three exposure blocks, and then created three Latin-square designed lists that counter-balanced the order of these blocks across participants.

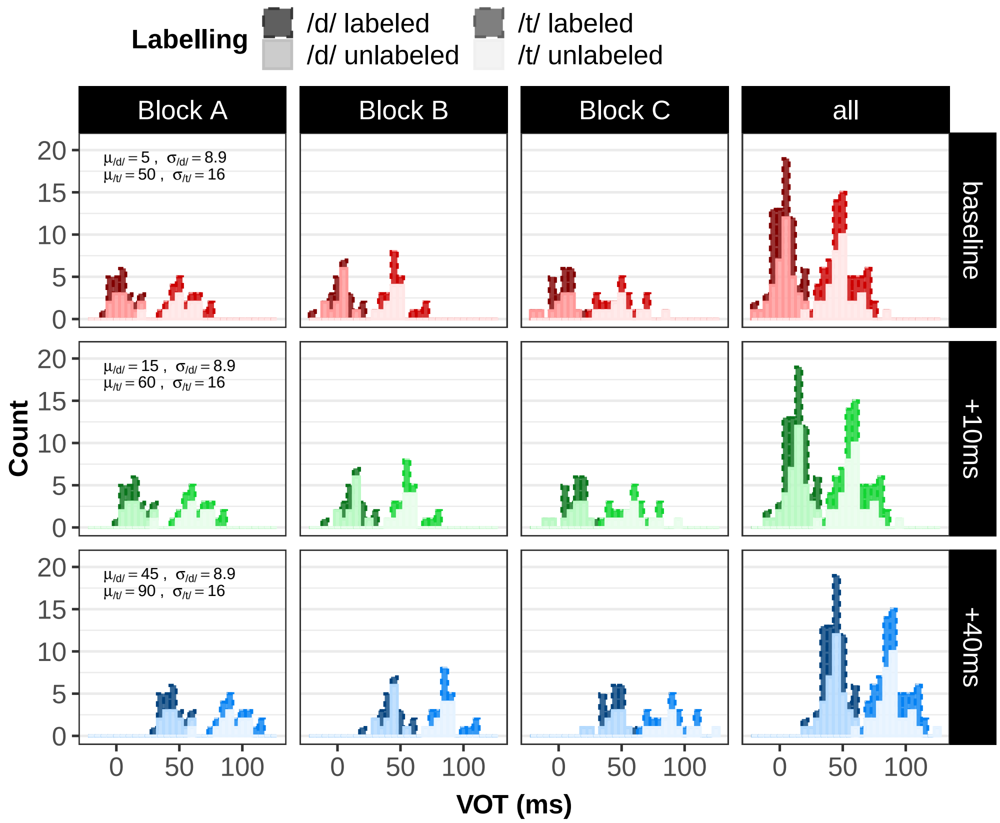

Figure 3. Histogram of VOTs for each of the three exposure blocks A-C by trial type (/d/ or /t/, labeled or unlabeled) and exposure condition (baseline vs. +10 vs. +40). Each exposure block contained 12 labeled /d/, 12 labeled /t/, 12 unlabeled /d/, and 12 unlabeled /t/ trials, as well as 6 catch trials (not shown). Except for the shift in VOTs (+0, 10 or 40 ms VOT to each trial), the VOT distribution of these trials–as well as the relative placement of labeled and unlabeled trials—was identical across exposure conditions. The order of exposure blocks A-C was counter-balanced across participants within each exposure condition using a Latin-square design.

Half of the /d/ and half of the /t/ trials in each exposure block were labeled, the other half was unlabeled. Earlier distributional learning studies have mostly used fully unlabeled exposure (Bejjanki et al., 2011; Clayards et al., 2008; Nixon et al., 2016). This contrasts with visually- or lexically-guided perceptual learning studies, which use labeled exposure (Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris et al., 2003; Vroomen et al., 2007). Such labeling is known to facilitate adaptation (Burchill, Liu, & Jaeger, 2018; **burchill2023?**; but see Kleinschmidt et al., 2015)—indeed, if shifted pronunciations are embedded in minimal pair or nonce-word contexts, listeners do not shift their categorization boundary (Norris et al., 2003; **REF-theodore?**; **babel?**). While lexical contexts often disambiguate sounds in everyday speech, that is not *always* the case: especially, when confronted with unfamiliar accents, listeners often have uncertainty about the word they are hearing, and must either use contextual information to label the input or adapt from unlabeled input. Here, we thus aimed to strike a compromise between always and never labeling the input (following one of the conditions in Kleinschmidt et al., 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. 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 all VOTs sampled for the baseline condition by +10 or +40 ms (see Figure 3). 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 exactly constant across conditions except for the shift in VOT—including the placement of labeled and unlabeled trials relative to the exposure condition's category means. 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 18 lists, obtained by crossing 3 (exposure condition) x 3 (block order) x 2 (placement of response options during unlabeled test and exposure trials).

2.4 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 = 0), participants who had atypical categorization 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), approximately evenly split across the three exposure conditions.

3 Results

We analyzed participants' categorization responses during exposure and test blocks in two separate Bayesian mixed-effects psychometric models, using brms (Bürkner, 2017) in R (R Core Team, 2022; RStudio Team, 2020).[3] Psychometric models account for attentional lapses while estimating participants' categorization functions. Failing to account for attentional lapses—while commonplace in research on speech perception (but see Clayards et al., 2008; Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (Prins, 2011; Wichmann & Hill, 2001). For the present experiment, lapse rates were negligible (0.8%, 95%-CI: 0.4 to 1.5%), and all results replicate in simple mixed-effects logistic regressions (Jaeger, 2008). This lapse rate compares favorably against those assumed or reported in prior work (Clayards et al., 2008; Kleinschmidt, 2020; e.g., Kleinschmidt & Jaeger, 2016).

The psychometric models for exposure and test blocks each regressed participants' categorization responses against the full factorial interaction of VOT, block, and exposure condition, along with the maximal random effect structure (by-subject intercepts and slopes for VOT, block, and their interaction, and by-item intercept and slopes for the full factorial design; see SI, **??**). All hypothesis tests reported below are based on these models. Figure 4 summarizes the results that we describe in more detail next. Panels A and B show participants' categorization responses during exposure and test blocks, along with the categorization 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 categorization function across blocks and conditions. These panels facilitate comparison across blocks within each exposure condition. Here we focus on the test blocks, which

[3] Fitting the models separately avoids questions about how differences in the VOT distribution during exposure blocks might affect the analysis of test blocks. For the test analyses, it also removes any potential collinearity between effects of exposure and effects of VOT.

were identical within and across exposure conditions.[4] 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 analyzed the *average* effect of exposure across the entire experiment ('batch tests,' e.g., Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; Nixon et al., 2016; Theodore & Monto, 2019) or during a single post-exposure test (e.g., Kleinschmidt, 2020). Then we present novel analyses that address questions about the incremental adaptation—testing the predictions described in the introduction.

```
## [1] "VOT test mean: 35.8333333333333"
```

```
## [1] "VOT test mean: 35.8333333333333"
```

```
## [1] "VOT test mean: 35.8333333333333"
## [1] "VOT test mean: 35.8333333333333"
```

[4] Previous studies have estimated changes in participants' categorization responses by analyzing responses on unlabeled exposure trials (e.g., Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; Theodore & Monto, 2019). This approach compares responses across different VOT values (since the exposure inputs differed by exposure condition), increasing the risk that assumptions baked into the analysis approach (e.g., linearity along the acoustic-phonetic continuum) bias the results. The analysis of test blocks that are identical within and across participants avoids this issue.

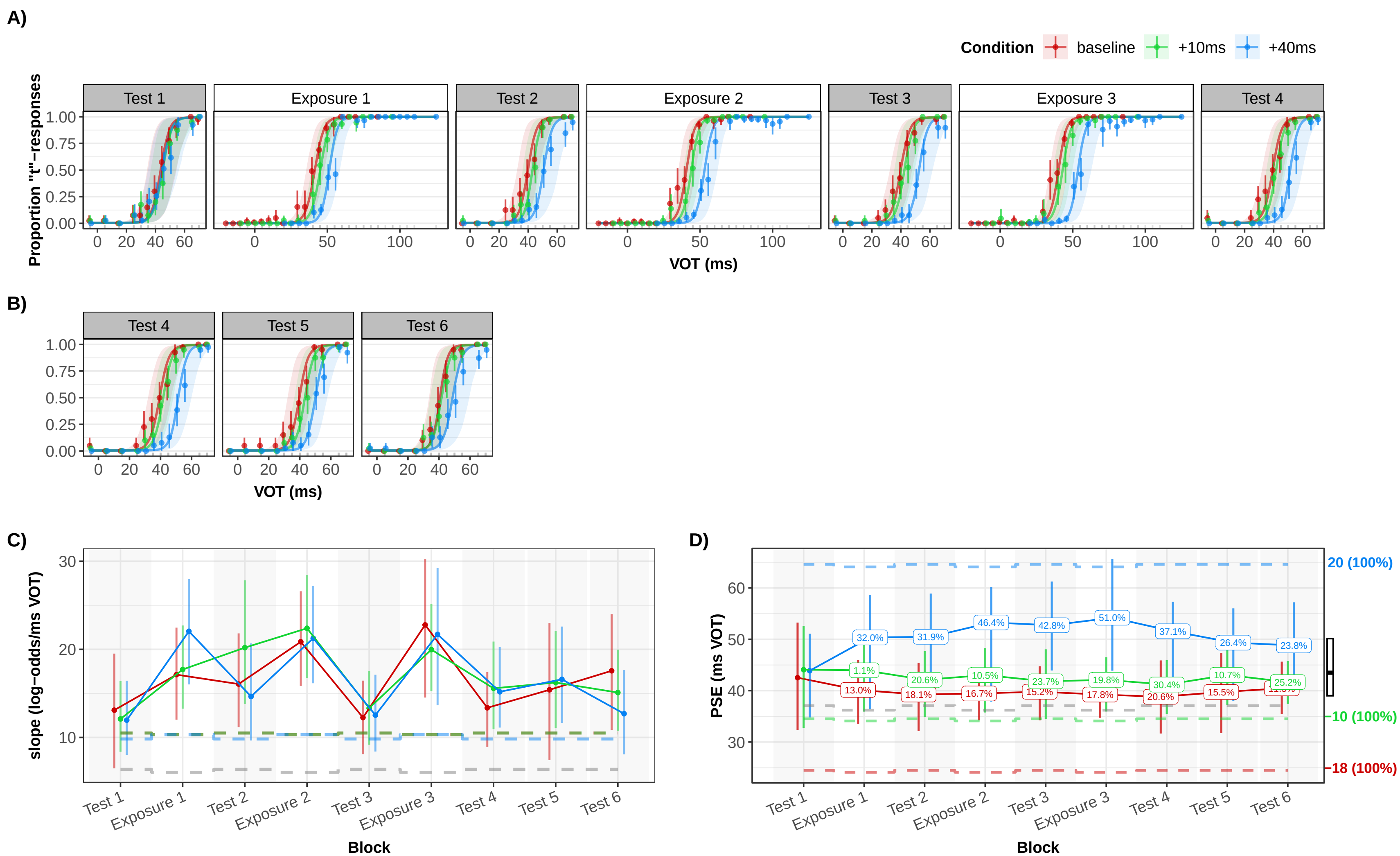

Figure 4. Summary of results. **Panel A:** Changes in listeners psychometric categorization functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks (only unlabeled trials were included in the analysis of exposure blocks since labeled trials provide no information about listeners' categorization function). Point ranges indicate the mean proportion of participants' "t"-responses and their 95% bootstrapped CI. Lines and shaded intervals show the *maximum a posteriori* (MAP) estimates 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 categorization functions shown in Panels A & B. Point ranges represent the posterior medians and their 95% CI. Dashed reference lines show the intercepts and PSEs that naive learner would be expected to converge against after sufficient exposure (an ideal observer model that has fully learned the exposure distributions). Percentage labels

3.1 Replication of previous findings (comparing exposure conditions averaging over test 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 longer the VOT ($\hat{\beta} = 15.09,\ 90\%{-}\mathrm{CI} = [12.377, 17.625],\ BF \geq 8000,\ p_{posterior} = 1$). Critically, exposure affected participants' categorization responses in the expected direction. Marginalizing over all test blocks, participants in the +40 condition were less likely to respond "t" than participants in the +10 condition ($\hat{\beta} = -2.26,\ 90\%{-}\mathrm{CI} = [-3.258, -1.228],\ BF = 162.3,\ p_{posterior} = 0.994$) or the baseline condition ($\hat{\beta} = -3.08,\ 90\%{-}\mathrm{CI} = [-4.403, -1.669],\ BF = 215.2,\ p_{posterior} = 0.995$). 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 ($\hat{\beta} = -0.82,\ 90\%{-}\mathrm{CI} = [-1.887, 0.282],\ BF = 8.9,\ p_{posterior} = 0.899$). That is, the +10 and +40 conditions resulted in categorization functions that were shifted rightwards compared to the baseline condition, as also evident in Figures 4.

This replicates previous findings that exposure to changed VOT distributions changes listeners' categorization responses (for /b/-/p/: Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt et al., 2015; for /g/-/k/, Theodore & Monto, 2019). Next, we turn to the questions of primary interest. Incremental changes in participants' categorization responses can be assessed from three mutually complementing perspectives. First, we compare how exposure affects listeners' categorization responses *relative to other exposure conditions*. This tests how early in the experiment differences between exposure conditions begin to emerge. Second, we compare how exposure changes listeners' categorization responses from block to block within each condition, relative to listeners' responses prior to any exposure. Third and finally, we compare changes in listeners' responses to those expected from an ideal observer that has fully learned the exposure distributions. This analysis can identify constraints on cumulative adaptation. For all three analyses, we initially focus on Tests 1-4 with intermittent exposure. Following that, we analyze the consequences of repeated testing during Tests 4-6, which have methodological relevance for future work.

3.2 How quickly does exposure affect listeners' categorization responses? (comparing exposure conditions within each block)

Figure 4A suggests that differences between exposure conditions emerged early in the experiment: already in Test 2, listeners in the +10 condition have shifted their categorization functions rightwards relative to the baseline condition, and listeners in the +40 condition have shifted their in categorization functions even further rightwards. This is confirmed by Bayesian hypothesis tests summarized in Table 1. Prior to any exposure, during Test 1, participants' responses did not differ across exposure condition. This result is predicted by models of adaptive speech perception under the assumptions that (a) participants in the different groups have similar prior experiences, and that (b) our sample size of is sufficiently large to yield stable estimates of listeners' categorization function.

During Test 2, after exposure to only 24 /d/ and 24 /t/ stimuli (thereof half labeled), participants' categorization responses already differed between exposure conditions (BFs > 13.7). The differences between exposure conditions that emerged at this point were all in the direction predicted by models of adaptive speech perception. Additional analyses reported in the SI (**??**) found that listeners' categorization functions had already changed *during* the first exposure block, in line with Figure 4A. This suggests that changes in listeners' categorization responses emerged *quickly* at the earliest point tested—after only a fraction of exposure trials previously tested in similar paradigms.

The effects of the three exposure conditions continued to persist until Test 4. Table 1 does, however, indicate an interesting non-monotonic development in the way that listeners' categorization function changed. While the difference between the +40 condition and both the baseline and +0 condition continued to increase numerically with increasing exposure (increasingly larger magnitude of negative estimates in Tests 2-4), the same was not the case for the difference between the +10 and the baseline condition. Instead, the difference between the +10 and baseline condition reduced with increasing exposure (while maintaining its direction). This development turns out to be potentially important in understanding incremental adaptation, and we continue to discuss it below.

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. Note that righward shifts correspond to negative effects (lower intercepts in predicting the log-odds of "t"-responses).

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Test block 1 (pre-exposure)					
+10 vs. baseline = 0	-0.34	0.75	[-2.025, 1.437]	3.3	0.77
+40 vs. +10 = 0	0.25	0.73	[-1.338, 1.903]	3.7	0.79
+40 vs. baseline = 0	-0.08	0.91	[-2.124, 2.082]	4.8	0.83
Test block 2					
+10 vs. baseline	-1.45	0.88	[-2.933, 0.181]	13.7	0.93
+40 vs. +10	-2.08	0.99	[-3.824, -0.173]	24.3	0.96
+40 vs. baseline	-3.49	1.24	[-5.635, -1.072]	54.2	0.98
Test block 3					
+10 vs. baseline	-0.78	0.62	[-1.888, 0.364]	7.9	0.89
+40 vs. +10	-2.80	0.82	[-4.188, -1.113]	86.0	0.99
+40 vs. baseline	-3.56	0.97	[-5.202, -1.582]	110.1	0.99
Test block 4					
+10 vs. baseline	-0.88	0.85	[-2.36, 0.847]	4.8	0.83
+40 vs. +10	-3.32	0.89	[-4.883, -1.636]	128.0	0.99
+40 vs. baseline	-4.16	1.21	[-6.275, -1.882]	122.1	0.99
Test block 5 (repeated testing without additional exposure)					
+10 vs. baseline	-1.33	0.71	[-2.556, -0.003]	19.1	0.95
+40 vs. +10	-2.38	0.86	[-3.893, -0.796]	65.1	0.98
+40 vs. baseline	-3.25	1.24	[-5.307, -0.923]	53.0	0.98
Test block 6 (repeated testing without additional exposure)					
+10 vs. baseline	-0.22	0.72	[-1.485, 1.114]	1.6	0.62
+40 vs. +10	-1.70	0.79	[-3.078, -0.171]	25.0	0.96
+40 vs. baseline	-2.57	1.22	[-4.58, -0.191]	24.0	0.96

3.3 Incremental adaptation from prior expectations (comparing block-to-block changes within exposure conditions)

Next, we compare how exposure affected listeners' categorization responses from block to block *within* each exposure condition. To facilitate visual comparison, Figure 4C & D summarize these changes for the slope and PSE, respectively. Focusing for now on Tests 1-4, this highlights four aspects of participants' behavior that were not readily apparent in the statistical comparisons we have summarized so far.

First, Panel C highlights the relative lack of changes in the slope of listeners categorization function. Slope changes, or lack thereof, have received comparatively attention in previous work (but see Clayards et al., 2008; Theodore & Monto, 2019) but they form part of the empirical facts that theories of speech perception need to account for. Compared to the changes in PSEs in Panel D, changes in the slope of listeners' categorization functions in Panel C were similar across exposure conditions (BFs < XXX; SI, **??**). Indeed, slopes changed very little relative to listeners' categorization responses in Test 1 (BFs < XXX; see SI, **??**). Both of these findings are in line with distributional learning theories of adaptive speech perception (Kleinschmidt & Jaeger, 2015), given that the variance of /d/ and /t/ was (a) held constant across all three exposure conditions, and (b) designed to resemble the variance of /d/ and /t/ in typical speech input.

Table 2
When did exposure begin to affect participants' categorization responses? This table summarizes the simple effects of block for each exposure condition. Note that righward shifts correspond to negative effects (lower intercepts in predicting the log-odds of "t"-responses).

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
baseline					
Test 2 vs. Test 1 > 0	1.17	0.71	[-0.218, 2.518]	12.87	0.93
Test 3 vs. Test 2 > 0	0.12	0.70	[-1.314, 1.477]	1.32	0.57
Test 4 vs. Test 3 > 0	0.16	0.54	[-0.863, 1.123]	1.72	0.63
Test 5 vs. Test 4 < 0	-0.36	0.49	[-1.275, 0.528]	3.52	0.78
Test 6 vs. Test 5 < 0	-0.57	0.61	[-1.655, 0.623]	4.63	0.82
Test 4 vs. Test 1 > 0	1.48	1.13	[-0.729, 3.441]	7.62	0.88
+10					
Test 2 vs. Test 1 > 0	0.16	0.79	[-1.168, 1.617]	1.42	0.59
Test 3 vs. Test 2 > 0	0.60	0.66	[-0.567, 1.85]	4.47	0.82
Test 4 vs. Test 3 > 0	0.17	0.77	[-1.324, 1.644]	1.40	0.58
Test 5 vs. Test 4 < 0	-0.58	0.58	[-1.626, 0.517]	4.88	0.83
Test 6 vs. Test 5 < 0	0.44	0.65	[-0.79, 1.651]	0.31	0.24
Test 4 vs. Test 1 > 0	0.94	1.21	[-1.305, 3.169]	3.46	0.78
+40					
Test 2 vs. Test 1 < 0	-2.06	0.79	[-3.428, -0.563]	45.24	0.98
Test 3 vs. Test 2 < 0	-0.73	0.78	[-2.093, 0.629]	4.74	0.83
Test 4 vs. Test 3 < 0	-0.06	0.81	[-1.48, 1.335]	1.11	0.53
Test 5 vs. Test 4 > 0	0.61	0.77	[-0.755, 1.928]	3.55	0.78
Test 6 vs. Test 5 > 0	0.75	0.72	[-0.56, 2.005]	5.55	0.85
Test 4 vs. Test 1 < 0	-2.86	1.12	[-4.868, -0.733]	50.28	0.98

Second, 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 2. To understand this pattern, it is helpful to relate the three exposure conditions to the distribution of VOT in listeners' prior experience. Figure 5 shows the category means of our exposure conditions relative to the distribution of VOT by talkers of L1-US English (based on Chodroff & Wilson, 2018). 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, only the +40 condition presented larger-than-expected category means, whereas the baseline condition and, to some extent, the +10 condition presented lower-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 (Table 2), and the direction of differences *between* exposure conditions receive an explanation (Table 1).

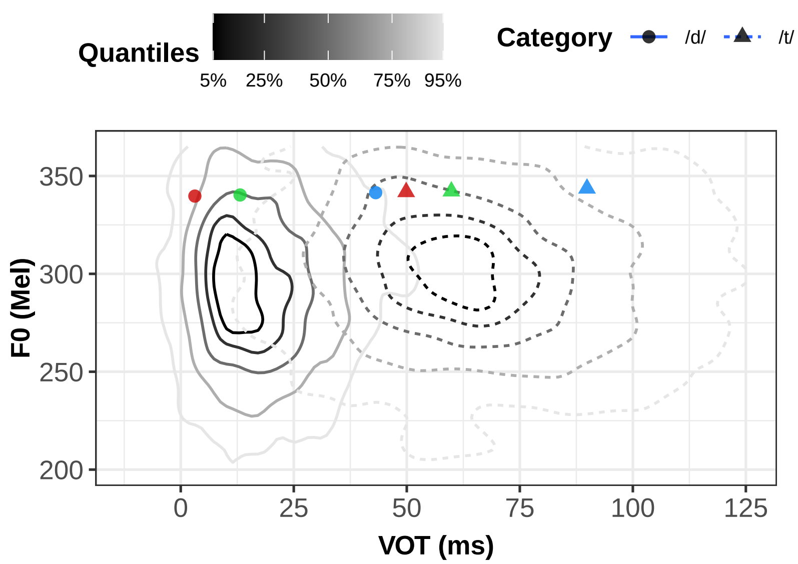

Figure 5. Placement of exposure stimuli relative to an estimate of typical phonetic distributions for 4,384 word-initial /d/ and /t/ productions in L1-US English [based on 72 female talkers in Chodroff and Wilson (2018); for details, see SI **??**]. The outermost contour of each category shows the 95% density quantile. Points show the category means of the exposure condition.

Third, the estimates in Table 2 suggest that listeners' PSEs changed most from Test 1 to Test 2, and then changed less and less with additional exposure up to Test 4 (smaller magnitude of estimates compared to earlier test blocks). This pattern is predicted by models of adaptive speech perception that are sensitive to the prediction error experienced while processing speech. This includes models that assume error-based learning (Sohoglu & Davis, 2016; Harmon et al., 2019; see also discussion in **davis-sohoglu2020?**) as well as Bayesian belief-updating models (Kleinschmidt & Jaeger, 2015; for demonstration, see **jaeger2019?**).

Fourth, Panel D also begins to illuminate the reasons for the non-monotonic development of the +10 and baseline conditions relative to each other, discussed in the previous section. In particular, this non-monotonicity does *not* appear due to a reversal of the effects in either of the two exposure conditions. Rather, both exposure conditions continue to change listeners' categorization function in the same direction from Test 1 to Test 4. However, after the rapid change from the pre-exposure Test 1 to the first post-exposure Test 2, listeners' categorization responses in the baseline condition did not change as much as in the +10 condition. Additional Bayesian hypothesis tests reported in the SI (**??**) suggest that these differences in the incremental effects of the two conditions are credible (BF = XXX). This explains the reduction in the difference between the +10 and baseline conditions discussed in the previous section. It does, however, raise the question *why* listeners' responses in the baseline condition did not change further with increasing exposure. Our third and final perspective on the incremental changes induced by exposure begins to address this question.

3.4 Constraints on cumulative adaptation (comparing exposure effects against idealized learner models)

Figures 4C & D also compare participants' responses against those of an idealized learner that has fully learned the exposure distributions. Specifically, we fit Bayesian ideal observers against the labeled VOT distributions of each exposure condition. The dashed lines represent the slopes and PSEs, respectively, that are expected from these models (for details, see SI **??**). This makes it possible to assess whether—or how much—listeners have converged against the exposure distributions. We make two observations.

First, the slopes of listeners' categorization functions in Panel C approximate those predicted by the idealized learner models: many of the 95% CIs overlap with the dashed lines.

This result, too, is predicted by distributional learning models of adaptive speech perception.[5]

Second, Panel D suggests that listeners did *not* converge against the exposure distributions. The percentage labels in Panel D quantify the degree to which listeners adapted their PSE towards the statistics of the exposure condition: 0% would correspond to no change relative to the listeners' PSE in Test 1, and 100% would correspond to complete convergence against the PSE predicted for an idealized learner. This highlights a striking asymmetry between the condition resulting in rightward shifts of the categorization function (+40), and the conditions resulting in leftward shifts (baseline and +10). On the one hand, the predicted PSEs of an idealized learner for the +40 and baseline conditions are shifted approximately by about the same amount relative to listeners' pre-exposure PSE in Test 1. However, the degree to which listeners converged against these predicted PSEs differed substantially between the two conditions, with cumulative adaptation proceeding almost twice as far in the rightward-shifted +40 condition (in Test 4: 37.1% towards idealized PSE) compared to the leftward-shifted baseline condition (20.6%). Comparing within just the leftward-shifted conditions, we find that relative shift is smaller for the baseline condition, compared to the +10 condition (30.4%).

3.5 Effects of repeated testing over the same uniform test continuum

Finally, we briefly summarize the effects of repeated testing evident in Tables 1 and 2. Some models of adaptive perception predict that exposure to uniformly distributed test tokens will reduce the effect of preceding exposure (Kleinschmidt & Jaeger, 2015; for relevant discussion, see also **winter-lancia2013?**). In line with these theories, we find that the effects of exposure reduced from Test 4 to Test 6. In Table 2, this is evident in a reversal of the direction of the block-to-block changes for Tests 5-6, compared to Tests 1-4. For the +40 exposure condition, these block to block changes went from rightward shifts in Tests 1-4 to leftward shifts in Tests 5-6. For the other two exposure conditions, the opposite pattern from leftward to rightward shift

[5] Of note, we followed Xie et al. (2023) and included perceptual noise in the ideal observer (estimated for VOT in Kronrod, Coppess, & Feldman, 2016). This deviates from some earlier comparisons of human perception against ideal observers (Clayards et al., 2008). Without the inclusion of perceptual noise, ideal observers predict much steeper categorization functions (offering a potential explanation for the mismatch between the ideal observer predictions and human categorization responses observed in Clayards et al., 2008). This highlights the importance of considering perceptual noise when modeling human speech perception (see also **burchill2023?**; **chodroff2016?**; **feldman2009?**).

was observed. As a consequence, exposure effects were substantially smaller in Test 6 than in Test 4 (see Table 1: while the effects of the +40 condition relative to the other two exposure conditions were still credible even in Test 6 (BFs > 24), this was no longer the case for the effect of the +10 condition relative to the baseline condition (BF = 1.6). This pattern of results replicates previous findings from LGPL (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019; Tzeng et al., 2021), and extends them to distributional learning paradigms (see also 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 & Jaeger, 2018). Analyses that average across all test tokens—as remains the norm—are bound to systematically underestimate the adaptivity of human speech perception.[6]

4 General discussion

- discuss rapid adaptation. link to findings from LGPL and VGPL [cummings-theodore; lj18,19]
- discuss fast-then-slow adaptation. link to findings in VGPL [kj11, 12, K20]
- discuss other evidence for constraints in DL work [kj16; k20], potentially also limits in vroomen 07, kj12 though these are harder to compare.
- discuss the fact that changes from block to block were largest at the beginning is consistent with the predictions of error-based learning (Sohoglu & Davis, 2016) and Bayesian inference (Kleinschmidt & Jaeger, 2015; for demonstration, see **jaeger2019?**).
- discuss consequences of findings for other accounts (decision-making; normalization)

[6] (**kraljic-samuel2006?**) is sometimes cited as finding LGPL exposure effects even after 480 test trials over a uniform test continuum. This is, however, misleading. Kraljic and Samuel used four *different* uniform test continua over two different phonetic contrasts (/b/-/p/ and /d/-/t/). Each test session consisted of 10 randomized repetitions of 6 test trials. Kraljic and Samuel never tested (or made any claims about) whether exposure effects were still detectable during the 10th repetition. Rather they report *average* effects across the 10 repetitions (like other LGPL studies), which is perfectly compatible with the hypothesis that repeated testing reduces the effects of exposure (see Liu & Jaeger, 2018).

- 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. future studies could rerun the exact same paradigm but only test at position x (i.e., a between-subject version of our design)

- 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.

5 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.

Assmann, P. F., & Nearey, T. M. (2007). Relationship between fundamental and formant frequencies in voice preference. *The Journal of the Acoustical Society of America*, *122*(2), EL35–EL43.

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.

Bent, T., & Baese-Berk, M. (2021). Perceptual learning of accented speech. *The Handbook of Speech Perception*, 428–464.

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer. Version 6.2. 12.*

Bradlow, A. R., Bassard, A. M., & Paller, K. A. (2023). Generalized perceptual adaptation to second-language speech: Variability, similarity, and intelligibility. *The Journal of the Acoustical Society of America*, *154*(3), 1601–1613.

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

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PloS One*, *13*(8), e0199358.

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

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.

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

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

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88.

Hitczenko, K., & Feldman, N. H. (2016). Modeling adaptation to a novel accent. *Proceedings of the Annual Conference of the Cognitive Science Society.*

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

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*, 1744–1762.

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.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–146). San Diego: Academic Press.

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

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

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

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

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

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

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.

Lancia, L., & Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Laboratory Phonology*, *4*(1), 221–257.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

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*(12), 1562.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., et al.others. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), e12823.

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.

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing* ({PhD} dissertation). The University of Iowa.

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.

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

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

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

RStudio Team. (2020). *RStudio: Integrated development environment for r.* Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*, 355–367.

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

Schuster, S. (2020). *Praat: Doing phonetics by computer [computer program].* Stanford, CA: Interactive Language Processing Lab Stanford. Retrieved from https://docs.proliferate.alps.science/en/latest/contents.html

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

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756.

Tan, M., Xie, X., & Jaeger, T. F. (2021). Using rational models to interpret the results of experiments on accent adaptation. *Frontiers in Psychology*, 4523.

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.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*, *28*, 1003–1014.

Vroomen, J., Linden, S. van, De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577.

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., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619.

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*.

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

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