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

Maryann Tan[1,2], Maryann Tan[2,3], & T F Jaeger[2]

[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, YOUR ADDRESS. 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; perceptual adaptation; distributional learning; ...

Word count: X

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

TO-DO

0.1 Highest priority

- MARYANN

0.1.1 Priority

- FLORIAN

0.2 To do later

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

1 Introduction

Talkers vary in the way they realise linguistic categories. Yet, listeners who share a common language background typically cope with talker variability without difficulty. In scenarios where a talker produces those categories in an unexpected and unfamiliar way comprehension may become a real challenge. It has been shown, however that brief exposure to unfamiliar accents can be sufficient for the listener to overcome any initial comprehension difficulty (e.g. Bradlow & Bent, 2008; Clarke & Garrett, 2004; Xie, Liu, & Jaeger, 2021; Xie et al., 2018). This adaptive skill is in a sense, trivial for any expert language user but becomes complex when considered from the angle of acoustic-cue-to-linguistic-category mappings. Since talkers differ in countless ways and each listening occasion is different in circumstance, there is not a single set of cues that can be definitively mapped to each linguistic category. Listeners instead have to contend with many possible cue-to-category mappings and infer the intended category of the talker. How listeners achieve prompt and robust comprehension of speech in spite of this variability (the classic "lack of invariance" problem) remains the a longstanding question in speech perception research.

In the past two decades the hypothesis that listeners overcome the lack of invariance by learning the distributions of talkers' acoustic cue-to-linguistic category mappings has gained considerable influence in contemporary approaches to studying this problem. A growing number of studies have demonstrated that changes in listener behaviour through the course of a short experiment align qualitatively with the statistics of exposure stimuli (Clayards, Tanenhaus, Aslin, & Jacobs, 2008a; Cummings & Theodore, 2023 etc; D. F. Kleinschmidt & Jaeger, 2015, 2016; Theodore & Monto, 2019).

- For example when listeners are tasked with identifying word pairs like *beach-peach* contrasted by the voice-onset-time (VOT) cue they would exhibit categorisation behaviour that corresponds to the properties of the distributions from which these words are sampled. Listeners exposed to tokens from distribution with wide variances tend to have categorisation functions that are shallower than listeners who hear words sampled from distributions with narrow variances (Clayards et al. (2008a); Theodore and Monto (2019)). In such paradigms, the means of the categories are held constant usually at locations where listeners would expect. This is motivated by hypotheses that listeners implicit knowledge about spoken language

- THE AIM OF THIS STUDY- The study we report here builds on the pioneering work of Clayards et al. (2008a) and D. F. Kleinschmidt and Jaeger (2016) with the aim to shed more light on how listeners' initial interpretation of cues from a novel talker incrementally change as they receive progressively more informative input of her cue-to-category mappings.

POINTS-TO-MAKE

- Most of the work has focused on the outcome of exposure.
- Qualitatively, we know that exposing listeners to different distributions produces changes in categorisation behaviour towards the direction of the shifts.
- A stronger test for the computational framework is needed.
- The ideal adapter framework makes specific predictions about rational speech perception. For example, listeners' integrate the exposure with their prior knowledge and infer the cue-category distributions of a talker. Listeners hold implicit beliefs or expectations about the distributions of cues which they bring to an encounter.
- The strength of these beliefs has bearing on listener propensity to adapt to a new talker – the stronger the prior beliefs the longer it takes to adapt. Listeners' strengths in prior beliefs about the means and variances are represented by parameters in the computational model. Listener behaviour observed collectively, thus far which speaks to this framework of thinking should by now be able to indicate roughly what those parameter values are. But it looks like those parameters are biased by the length of exposure and the outcome during experiments. No one has confronted this issue of very quick but limited adaptation which can't be solved by giving more exposure trials.
- How do we distinguish the results from normalization accounts which can also explain adaptation but is not usually regarded as learning?

-[IMPROVING ECOLOGICAL VALIDITY OF PARADIGM] A secondary aim was to begin to address possible concerns of ecological validity of prior work. While no speech stimuli is ever ideal, previous work on which the current study is based did have limitations in one or two aspects:the artificiality of the stimuli or the artificiality of the distributions. For e.g. (Clayards et al., 2008a) and (D. F. Kleinschmidt & Jaeger, 2016) made use of synthesised stimuli that were robotic or did not sound human-like. The second way that those studies were limited was that the exposure distributions of the linguistic categories had identical variances (see also Theodore & Monto, 2019) unlike what is found in production data where the variance of the voiceless categories are typically wider than that of the voiced category (Chodroff & Wilson, 2017). We take modest steps to begin to improve the ecological validity of this study while balancing the need for control through lab experiments by employing more natural sounding stimuli as well as by setting the variances of our exposure distributions to better reflect empirical data on production (see section x.xx. of SI).

We designed the experiment to provide high statistical power to detect effects of exposure, both incrementally within each exposure condition, and cumulatively across exposure conditions. To this end, we employed the repeated exposure-test design shown in Figure 1. The use of test blocks that repeated same stimuli across blocks and exposure conditions deviates from previous work (Clayards, Tanenhaus, Aslin, & Jacobs, 2008b; D. F. Kleinschmidt, 2020; **kleinschmidt-jaeger2016?**). 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). Since previous work has found that repeated testing over uniform test continua can reduce or undo the effects of informative exposure (**cummings202X?**), we kept test blocks short, each consisting of only 12 trials. The final 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. Finally, as we detail below, our design also allowed us to measure adaptation during exposure.

1.1 Methods

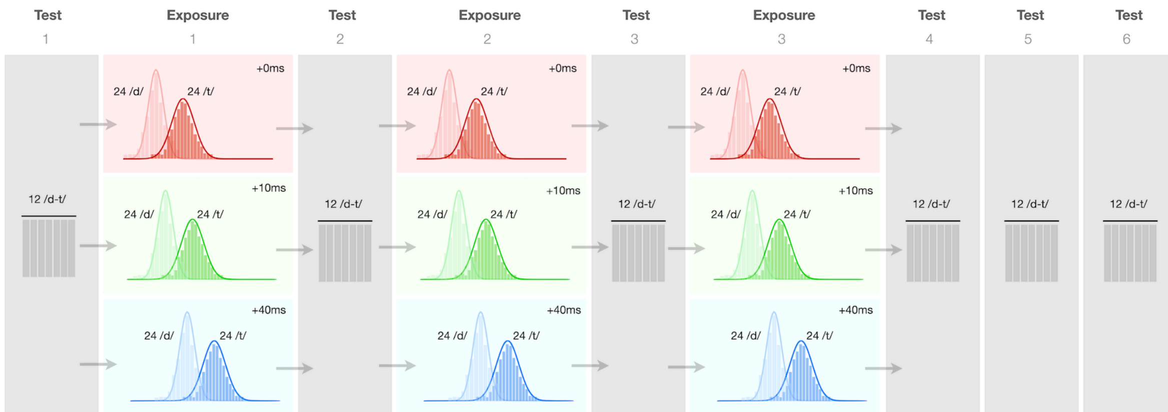

Figure 1. Exposure-test design of the experiment. Test blocks presented identical stimuli within and across conditions

1.1.1 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 only speak English, 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) spent at least the first 10 years of their life in the US speaking only English, (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 XXX participants loaded the experiment but did not start or complete it.

Participants took an average of 17 minutes to complete the experiment (SD = 9 minutes) –excluding time taken for instructions and survey– 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?**).

1.1.2 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. These recordings were used to create four natural-sounding minimal pair VOT continua using a script (Winn, 2020) in Praat (**praat?**). The VOTs generated for each continuum ranged from -100 to +120 msec in 5 msec steps.[1] 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, **??**).

This approach resulted in continuum steps that sound natural (unlike the highly robotic-sounding stimuli employed in Clayards et al., 2008a; D. F. Kleinschmidt & Jaeger, 2016). A post-experiment survey asked participants whether "XXX". No participant reported that the stimuli sounded unnatural (in contrast to other experiments we have conducted with robotic-sounding stimuli like those of **clayards?**). In addition to the critical minimal pair continua 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.

A norming experiment (N = 24 participants) reported in the SI (@**??**XXX)) was used to select the three minimal pairs that elicited the most similar categorization responses (dill-till, din-tin, and dip-tip). These three continua were used to create the three exposure conditions shown in Figure 1.

[1] For simplicity's sake, we follow previous work (D. F. 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, pevoicing is estimated to occur on XXX% of word-initial voiced stops and 0% of voiceless stops (**REF?**).

1.1.3 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 (**REF?**), and were instructed to not change the volume throughout the experiment. Following instructions, participants completed 234 two-alternative forced-choice categorisation trials (Figure **??**). 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.

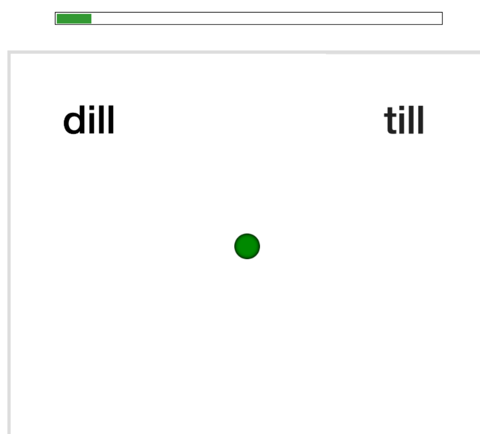

Figure 2. Example trial display. When the green button turned bright green, participants had to click on it to play the recording.

Unbeknownst to participants, the 234 trials were split into exposure blocks (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.

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 counter-balanced within and across test blocks, so that 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.

Both the placement of the response options (/d/ on the left vs. right) and the assignment of VOTs to minimal pair continua was counter-balanced across participants, using 2 x 3 Latin-square designed lists. Trial order was randomized within each block and participant.

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).

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 3). 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., 2008a; D. F. Kleinschmidt, 2020; **kleinschmidt-jaeger2016?**). We created one random sample for each of the three exposure blocks. Both the random seed and the order of exposure blocks was counter-balanced across participants using 3 (block order) Latin-squared designed exposure lists.

Half of the /d/ and half of the /t/ trials were labeled, the other half was unlabeled (paralleling one of the conditions in D. 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. 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 3). This approach exposes participants to heterogenous approximations of normally distributed VOTs for /d/ and /t/ that varied across blocks, while holding all aspects of the input constant across conditions except for the shift in VOT.

The order of trials was randomized within each block and participant, with the constraint that no more than two catch trials would occur in a row. Participants were randomly assigned to one of the 3 (exposure condition) x 3 (block order) x 2 (image mapping) exposure lists.

1.1.4 Exclusions

```
## Warning: There were 42 warnings in `mutate()`.
## The first warning was:
## i In argument: `CategorizationModel = map(...)`.
## i In group 2: `ParticipantID = 119`, `Experiment = AE-DLVOT`, `Condition.Exposure = Shift0`
## Caused by warning:
## ! glm.fit: fitted probabilities numerically 0 or 1 occurred
## i Run `dplyr::last_dplyr_warnings()` to see the 41 remaining warnings.

## Warning: Using one column matrices in `filter()` was deprecated in dplyr 1.1.0.
## i Please use one dimensional logical vectors instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

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

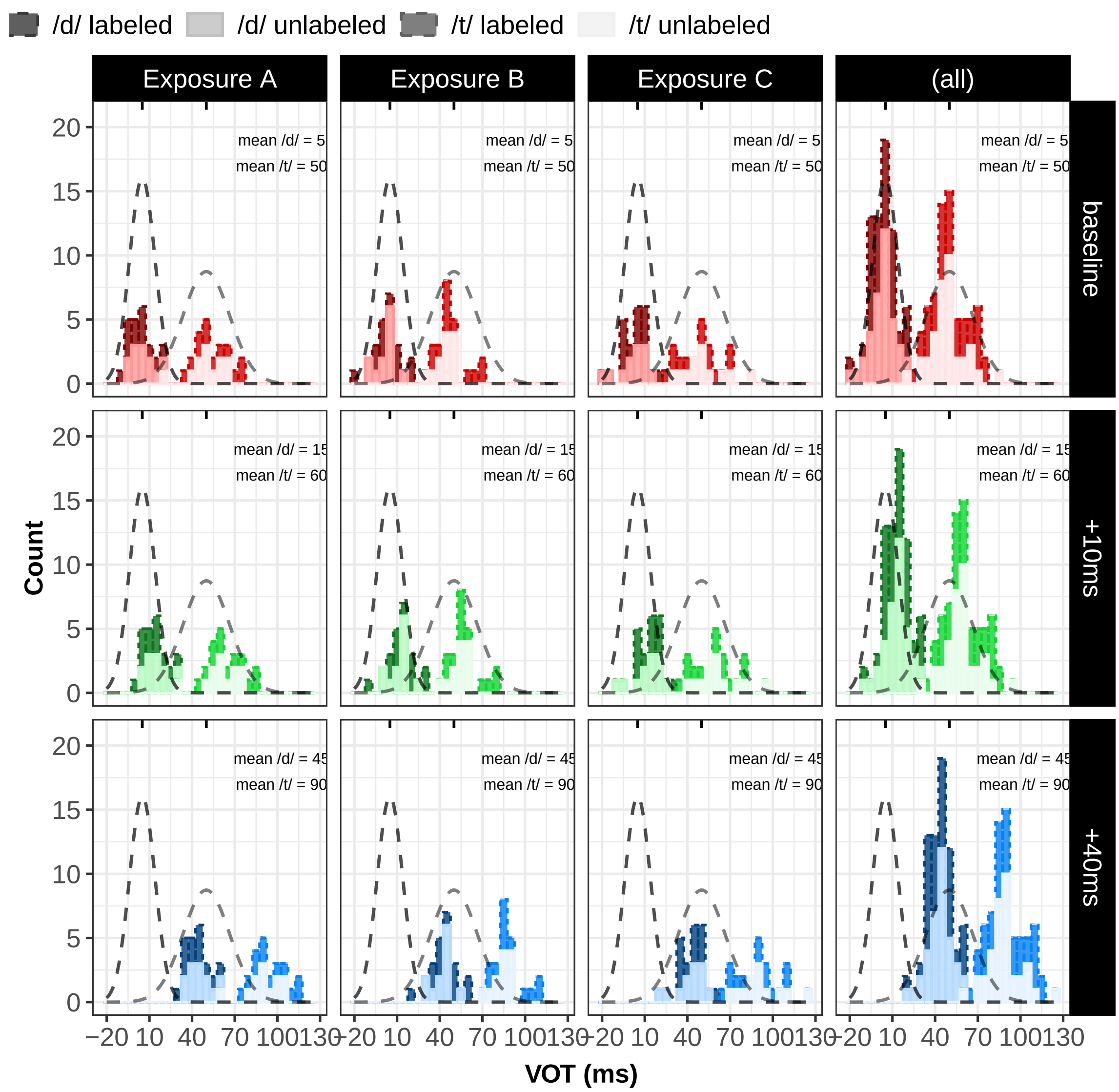

Figure 3. Histogram of VOTs across the 48 trials of all three exposure blocks by exposure condition. Shown is the parameterization of the VOT samples used across experimental lists. The order of blocks was counter-balanced across participants.

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 =) or not to be L1 speakers of US English (N = 0).

1.2 Results

1.3 Research questions and hypotheses

1. Do listeners change their categorization behaviour in the direction predicted by their respective exposure distributions?
2. At what stage in the experiment did the behavioural change first emerge?
3. Are the shifts in categorisation behaviour proportional to the differences between the exposure conditions?
4. Do the differences between exposure conditions diminish with repeated testing and without intermittent exposure?

[MORE HERE]

1.3.1 Regression analyses

Figures 4A-B summarize participants' categorisation responses during exposure and test blocks, depending on the exposure condition and VOT.

We analyzed participants' categorisation responses during test blocks in a Bayesian mixed-effects psychometric model (e.g., Prins, 2012). The psychometric model is an extension of mixed-effects logistic regression that also takes into account attentional lapses. Though we confirmed for the present case that all results would replicate in a simple mixed-effects logistic regression (Jaeger, 2008), ignoring attentional lapses—while commonplace in research on speech perception (but see Clayards et al., 2008a; D. F. Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (Wichmann & Hill, 2001).

The mixed-effects psychometric model describes the probability of “t”-responses as a weighted mixture of a perceptual and a lapsing model. The perceptual model predicts responses on trials where participants pay attention and respond based on the stimulus. We implemented the perceptual model as used mixed-effects logistic regression, predicting “t”-responses from exposure condition (backward difference coded, comparing the +10ms against the +0ms shift condition, and the +40ms against the +10ms shift condition), test block (backward difference coded from the first to last test block), VOT (Gelman scaled), and their full factorial interaction. The model included by-participant random intercepts and slopes for all within-participant maniplations (block and VOT) and by-item random intercepts and slopes for all within-participant manipulations (exposure condition, block, VOT).

The lapsing model predicts participant responses that are made independent of the stimulus—for example, responses that result from attentional lapses. These responses depend only on participants’ response bias. We used mixed-effects logistic regression with only a population-level intercept, allowing non-uniform responses bias but assuming that response biases did not vary across participants. Finally, the relative weight of the perceptual and lapsing model is determined by the lapse rate. We again used mixed-effects logistic regression with only a population-level intercept, inferring lapse rates from that data while assuming that lapse rates did not vary across participants or blocks (as confirmed by Figures **??**A-B).

We fit the psychometric model using the package `brms` (Bürkner, 2017) in R (R Core Team, 2021; RStudio Team, 2020). To faciltiate comparison of effect sizes across predictors, we standardized continuous predictors (VOT) by dividing through twice their standard deviation (Gelman, 2008). Following previous work from our lab (Hörberg & Jaeger, 2021; Xie et al., 2021), we used weakly regularizing priors to facilitate model convergence. For fixed effect parameters, we used Student priors centered around zero with a scale of 2.5 units (following Gelman, Jakulin, Pittau, & Su, 2008) and 3 degrees of freedom. For random effect standard deviations, we used a Cauchy prior with location 0 and scale 2, and for random effect correlations, we used an uninformative LKJ-Correlation prior with its only parameter set to 1, describing a uniform prior over correlation matrices (**Lewandowski2009?**). Four chains with 2000 warm-up samples and 2000 posterior samples each were fit. No divergent transitions after warm-up were observed, and all $1 < \hat{R} < 1.01$.

```
## [1] "VOT mean: 42.165"
## [1] "VOT sd: 30.3259"
## [1] "mean VOT is 42.1650326797386 and SD is 30.3259185098252"
##   _Exposure2 vs. Exposure1 _Exposure3 vs. Exposure2
## 2                    -0.67                    -0.33
## 4                     0.33                    -0.33
## 6                     0.33                     0.67
```

```
## [1] "VOT mean: 42.9636"
## [1] "VOT sd: 30.9118"
## [1] "mean VOT is 42.9636437908497 and SD is 30.9117519390561"
##   _Exposure2 vs. Exposure1 _Exposure3 vs. Exposure2
## 2                    -0.67                    -0.33
## 4                     0.33                    -0.33
## 6                     0.33                     0.67
```

```
## [1] "VOT mean: 42.165"
## [1] "VOT sd: 30.3259"
## [1] "mean VOT is 42.1650326797386 and SD is 30.3259185098252"
##   _Exposure2 vs. Exposure1 _Exposure3 vs. Exposure2
## 2                    -0.67                    -0.33
## 4                     0.33                    -0.33
## 6                     0.33                     0.67
## [1] "VOT mean: 35.8333"
## [1] "VOT sd: 22.1592"
## [1] "mean VOT is 35.8333333333333 and SD is 22.1591861746958"
##         _Shift10 vs. Shift0 _Shift40 vs. Shift10
## Shift0                -0.67                -0.33
## Shift10                0.33                -0.33
```

```
## Shift40                0.33                0.67
##   _Test2 vs. Test1 _Test3 vs. Test2 _Test4 vs. Test3 _Test5 vs. Test4 _Test6 vs. Test5
## 1 -5/6             -2/3             -1/2             -1/3             -1/6
## 3  1/6             -2/3             -1/2             -1/3             -1/6
## 5  1/6              1/3             -1/2             -1/3             -1/6
## 7  1/6              1/3              1/2             -1/3             -1/6
## 8  1/6              1/3              1/2              2/3             -1/6
## 9  1/6              1/3              1/2              2/3              5/6
```

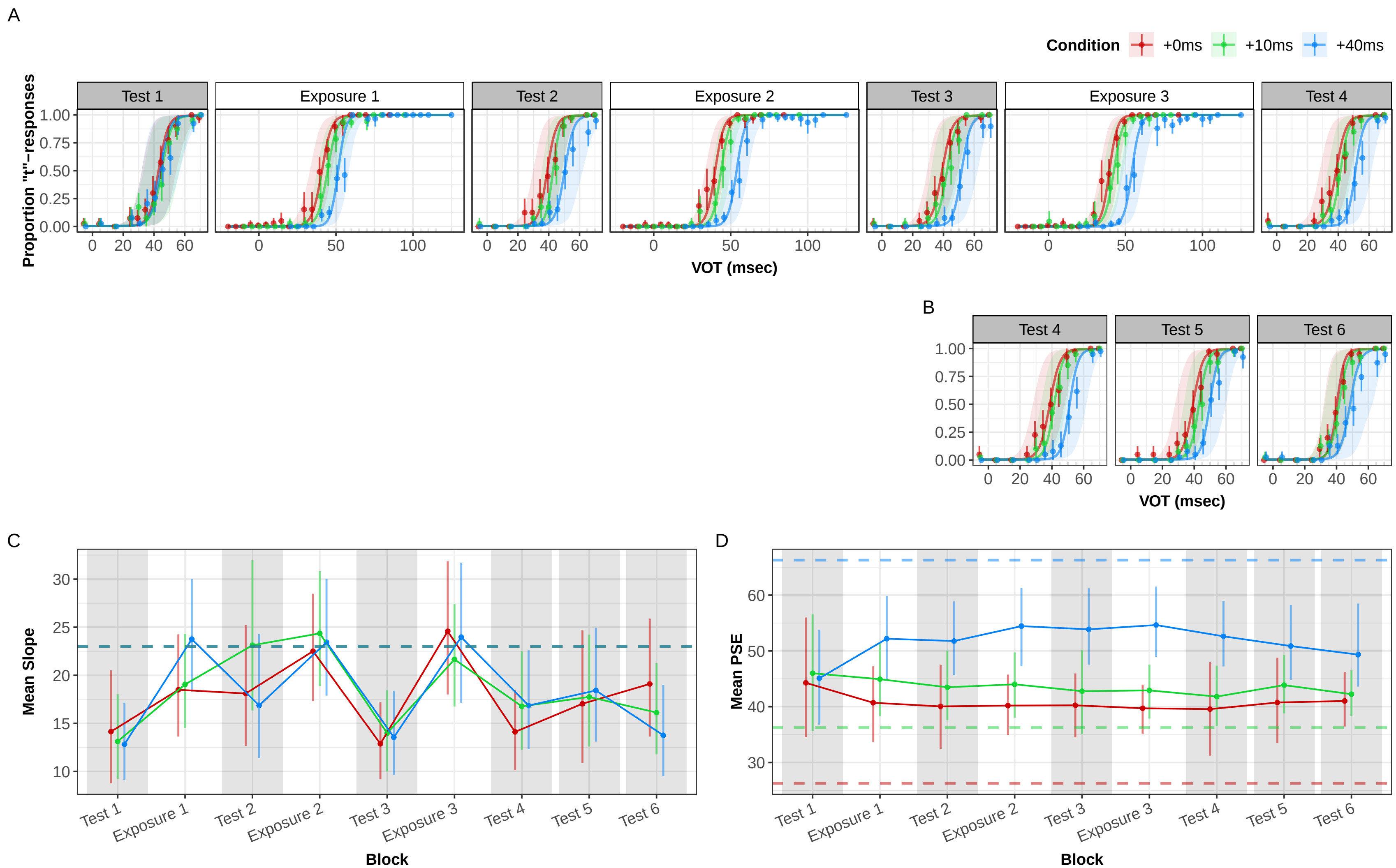

Figure 4. Summary of results. **Panel A:** Changes in listeners psychometric categorisation functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks. Point ranges indicate the mean proportion of "t"-responses and their 95% bootstrapped CI. Lines and shaded intervals 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).

1.4 Description of the overall pattern of results (main effects)

- The overall lapse rate was negligible ($\hat{\beta}$ = NA %, 95%-CI: NA to NA%; Bayes factor: Inf 90%-CI : -5.39 to -4.24) indicating that participants were paying attention in the majority of trials.
- There was a main effect of VOT ($\hat{\beta}$ = 15.7 95%-CI: 12.5 to 19.2; Bayes factor: 7,999 90%-CI : 13.15 to 18.4): participants were more likely to respond "t" as VOT increased.
- Condition had a main effect on responses such that with larger shifts away from the baseline, participants responded with fewer "t"s.
- Comparing the +10ms condition with the baseline condition across all blocks: there was a reduction in log-odds of responding "t" in the +10ms condition compared to the baseline condition ($\hat{\beta}$ = -1 95%-CI: -2.8 to 0.7; Bayes factor: 9.24 90%-CI : -2.24 to 0.3).
- Comparing the +40ms against the +10ms condition across all blocks: there was a reduction in log-odds of responding "t" in the +40ms condition compared to the +10ms condition ($\hat{\beta}$ = -2.4 95%-CI: -3.8 to -1.1; Bayes factor: 443.44 90%-CI : -3.54 to -1.36).
- Tellingly, the reduction in log-odds was larger in the +40 vs +10ms comparison, reflecting the larger magnitude of shift from the baseline (Bayes factor: 9.28 90%-CI : -3.36 to 0.44).

1.4.1 Interactions

The interactions provide between block comparisons of the differences between conditions. We focus on the first 4 test blocks as they were interspersed with exposure. In order to examine the effects of exposure condition on behaviour within block, and how each condition changed by block (simple effects of condition and block) we fitted 2 nested models that embed condition within block, and block within condition. We report the interactions in conjunction with the simple effects.

- Comparing the change in differences between +10ms and baseline between blocks: we see an overall reduction in the log-odds of responding "t" between test blocks 1 and 4 however almost all that reduction took place between test blocks 1 and 2 ($\hat{\beta}$ = -1.4 95%-CI: -3.5 to 0.6; Bayes factor: 13.52 90%-CI : -3.06 to 0.2). Between test blocks 2 and 4, differences in behaviour between the two groups did not change significantly in spite of increased input from the exposure blocks.

- Comparing the change in differences between +40ms and +10ms between blocks: -There was a consistent reduction in log-offs of responding "t" from blocks 1 through 4, indicating an incremental shift in categorisation towards the right as participants received more input. The biggest change was observed between test block 1 and 2 ($\hat{\beta} =$ -2.1 95%-CI: -4.4 to 0.2; Bayes factor: 27.78 90%-CI : -3.89 to -0.23).
  - The difference between condition +40 and +10 continued to widen after the second exposure block, ($\hat{\beta} =$ -1.8 95%-CI: -4.1 to 0.5; Bayes factor: 19.15 90%-CI : -3.69 to 0) but not much incremental shift was observed in the 4th test block in spite of full exposure to the 144 trials at this stage ($\hat{\beta} =$ -0.5 95%-CI: -3.3 to 2.1; Bayes factor: 1.69 90%-CI : -2.63 to 1.62)

```
## Warning in tidy.brmsfit(fit_mix_test_nested_block, effects = "fixed"): some parameter names
```

```
## Warning in tidy.brmsfit(fit_mix_test_nested_condition, effects = "fixed"): some parameter na
```

Table 1
Was there incremental change from test blocks 1 to 4?

Hypothesis	Estimate	Est Error	CI Lower	CI Upper	Evid Ratio	Post Prob
Difference in +10 vs baseline						
Test block 2 > Test block 1	-1.41	1.1	-3.1	0.20	13.52	0.93
Test block 3 > Test block 2	0.83	1.3	-1.1	2.78	0.25	0.20
Test block 4 > Test block 3	0.01	1.3	-1.8	1.89	1.02	0.50
Test block 4 > Test block 1	-0.57	1.9	-3.6	2.48	1.82	0.64
Difference in +40 vs +10						
Test block 2 > Test block 1	-2.06	1.2	-3.9	-0.23	27.78	0.96
Test block 3 > Test block 2	-1.81	1.2	-3.7	0.00	19.15	0.95
Test block 4 > Test block 3	-0.47	1.6	-2.6	1.62	1.70	0.63
Test block 4 > Test block 1	-4.35	1.9	-7.2	-1.72	101.56	0.99

All data and code for this article can be downloaded fromhttps://osf.io/q7gjp/. This article is written in R markdown, allowing readers to replicate our analyses with the press of a button using freely available software (R, R Core Team, 2021; RStudio Team, 2020), while changing any of the parameters of our models. Readers can revisit any of the assumptions we make—for example, by substituting alternative models of linguistic representations. The supplementary information (SI, **??**) lists the software/libraries required to compile this document. Beyond our immediate goals here, we hope that this can be helpful to researchers who are interested in developing more informative experimental designs, and to facilitate the interpretation of existing results (see also Tan, Xie, & Jaeger, 2021).

2 General discussion

2.1 Methodological advances that can move the field forward

An example of a subsection.

Table 2

When did change emerge? Are differences proportional?

Hypothesis	Estimate	Est Error	CI Lower	CI Upper	Evid Ratio	Post Prob
Test block 1						
+10 vs baseline	-0.38	1.14	-2.1	1.40	1.99	0.66
+40 vs +10	0.22	1.14	-1.4	1.85	0.68	0.40
+40 vs baseline	-0.16	1.45	-2.4	2.04	1.32	0.57
+40 vs baseline > 3x +10 vs baseline	1.36	3.81	-4.6	7.35	0.51	0.34
Test block 2						
+10 vs baseline	-2.15	1.38	-4.3	-0.11	22.12	0.96
+40 vs +10	-2.11	1.38	-4.3	0.07	17.35	0.95
+40 vs baseline	-2.49	1.74	-5.3	0.31	14.47	0.94
+40 vs baseline > 3x +10 vs baseline	-0.98	3.76	-7.0	4.78	1.59	0.61
Test block 3						
+10 vs baseline	-0.88	0.94	-2.2	0.42	7.98	0.89
+40 vs 10	-3.31	1.15	-5.2	-1.62	169.21	0.99
+40 vs baseline	-3.69	1.59	-6.2	-1.18	65.67	0.98
+40 vs baseline > 3x +10 vs baseline	-2.17	3.64	-7.8	3.37	3.01	0.75
Test block 4						
+10 vs baseline	-1.06	1.34	-3.0	0.95	5.46	0.84
+40 vs 10	-4.07	1.19	-6.0	-2.28	420.05	1.00
+40 vs baseline	-4.44	1.62	-7.1	-1.93	149.94	0.99
+40 vs baseline > 3x +10 vs baseline	-2.93	3.66	-8.8	2.69	4.53	0.82

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.

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

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

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in american english. *Journal of Phonetics*, *61*, 30–47.

Table 3
Effects of repeated testing (test blocks 4 to 6)

Hypothesis	Estimate	Est Error	CI Lower	CI Upper	Evid Ratio	Post Prob
Difference in +10 vs baseline						
Test block 5 < Test block 4	-0.34	1.20	-1.73	1.1	0.42	0.30
Test block 6 < Test block 5	1.27	0.97	-0.14	2.7	13.95	0.93
Test block 6 > Test block 4	0.93	1.44	-0.92	2.9	0.23	0.19
Difference in +40 vs +10						
Test block 5 < Test block 4	1.41	1.25	-0.54	3.3	8.66	0.90
Test block 6 < Test block 5	0.58	1.18	-1.27	2.3	2.79	0.74
Test block 6 > Test block 4	1.98	1.53	-0.42	4.3	0.08	0.08

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. (2008b). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

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

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

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Hörberg, T., & Jaeger, T. F. (2021). A rational model of incremental argument interpretation: The comprehension of swedish transitive clauses. *Frontiers in Psychology*, *12*, 674202.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and

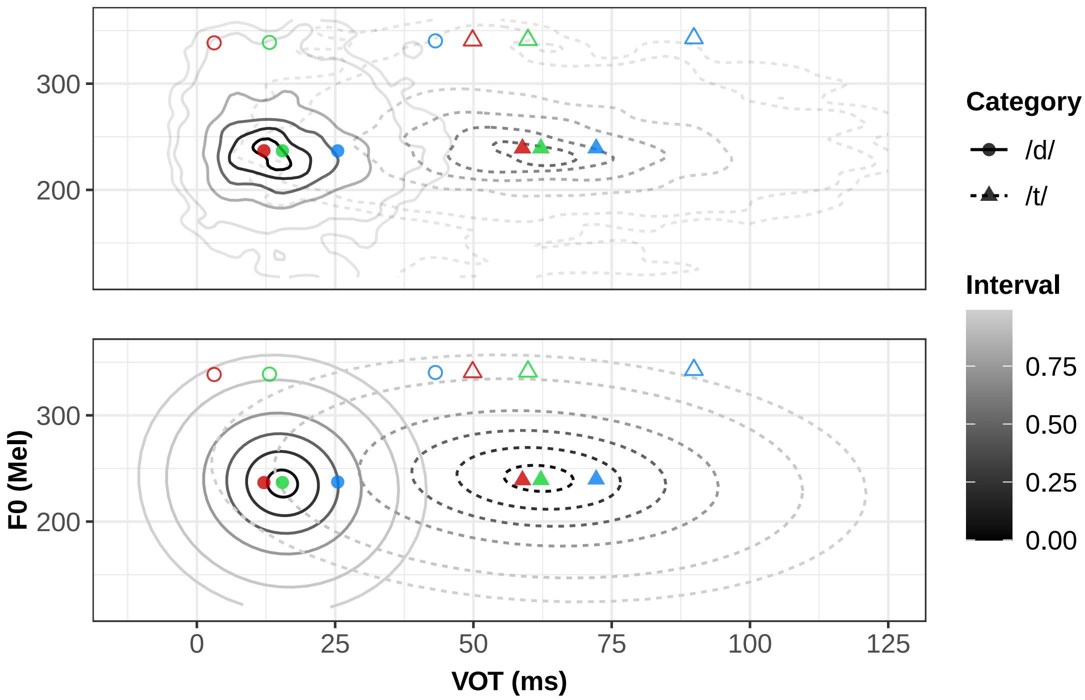

Figure 5

towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

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

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148. https://doi.org/https://doi.org/10.1037/a0038695

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

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

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology*.

Human Perception and Performance, *45*, 1562–1588. https://doi.org/10.1037/xhp0000693

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

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

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

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

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

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

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

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

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