Listeners adjust their prior expectations as they adapt to speech of an unfamiliar talker

Maryann Tan[1,2], T Florian Jaeger[2,3], & YOUR OTHER CO-AUTHOR[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

1 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

2 Listeners adjust their prior expectations as they adapt to speech of an unfamiliar talker

TO-DO

2.1 Highest priority

- MARYANN
  - Continue describing Experiment 2
  - Discuss with Florian for discussion
  - Fix any plot issues

2.1.1 Priority

- MARYANN
  - Fill in the references
- FLORIAN:
- Review Introduction
- Review Experiment 1 – comment on discussion of IO analysis
- Review plots
- Advise on how to adjust the text size of plot axis (theme() and element_text doesn't seem to work)

2.2 To do later

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

1 Introduction

Talkers who share a common language vary in the way they pronounce its linguistic categories. Yet, listeners of the same language background typically cope with such variation without much trouble. In scenarios where a talker produces those categories in an unexpected and unfamiliar way, comprehending their speech may pose a real challenge. However, brief exposure to the talker's accent (sometimes just minutes) can be sufficient for the listener to overcome any initial comprehension difficulty (e.g. Bradlow & Bent, 2008; Clarke & Garrett, 2004; X. Xie, Liu, & Jaeger, 2021; X. 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 accurate comprehension of speech in spite of this variability remains the overarching aim of speech perception research.

Researchers have been exploring the hypothesis that listeners solve this perceptual problem by exploiting their knowledge gained from experience with different talkers. This knowledge is often implicit and context contingent since listeners are sensitive to both social and environmental cues (e.g. age, sex, group identity, native language etc.) that are relevant for optimal speech perception. Impressively, shifts in perception can be induced implicitly through subtle cues such as the presence of cultural artefacts that hint at talker provenance, (Hay & Drager, 2010) and explicitly such as when the listener is instructed to imagine a talker as a man or a woman (Johnson, Strand, & D'Imperio, 1999). While these and other related effects of exposure-induced changes speak to the malleability of human perception, it remains unclear how human perceptual systems strike the balance between stability and flexibility.

One possibility is that listeners continuously update their implicit knowledge with each talker encounter by integrating prior knowledge of cue-to-category distributions with the statistics of the current talker's productions, leading to changes in representations which can be observed in listener categorisation behaviour. Broadly speaking, many theoretical accounts would agree with

this assertion. Connectionist (McClelland & Elman 1986; Luce & Pisoni, 1998), and Bayesian models of spoken word recognition (Norris & McQueen, 2008) and adaptation (D. F. Kleinschmidt & Jaeger, 2015) are generative systems that abstract the frequency of input. Even exemplar models of speech perception (Goldinger 1996, 1998; Johnson, 1997; Pierrehumbert 2001) which encode high fidelity memories of speaker-specific phonetic detail converge to a level of generalisation due to effects of token frequency (**Pierrehumbert2003?**; **DragerKirtley2016?**).

At the level of acoustic-phonetic input, listeners' implicit knowledge refer to the way relevant acoustic cues that distinguish phonological categories are distributed across talkers within a linguistic system. Talkers of US-English, for instance, distinguish the /d/-/t/ contrasts primarily through the voice-onset-time (VOT) acoustic cue. Given its relevance for telling word pairs such as "din" and "tin" apart, a distributional learning hypothesis would posit that listeners learn the distribution of VOT cues when talkers produce those stop consonant contrasts in word contexts. Earliest evidence for listener sensitivity to individual talker statistics in the domain of stop consonants come from studies such as Allen & Miller (2004, also **TheodoreMiller2010?**) but more recent studies that formalise the problem of speech perception as rational inference have shown that listeners' behavioural responses are probabilistic function of the exposure talker's statistics (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; D. F. Kleinschmidt & Jaeger, 2016; and Theodore & Monto, 2019).

Clayards et al. (2008) for instance found that listeners responded with greater uncertainty after they were exposed to VOT distributions for a "beach-peach" contrast that had wider variances as compared to another group who had heard the same contrasts with narrower variances. Across both wide and narrow conditions, the mean values of the voiced and voiceless categories were kept constant and set at values that were close to the expected means for /b/ and /p/ in US English. The study was one of the first to demonstrate that at least in the context of an experiment, listeners categorisation behaviour was a function of the variance of the exposure talker's cue distributions – listeners who were exposed to a wide distribution of VOTs showed greater uncertainty in their perception of the stimuli, exhibiting a flatter categorisation function on average, compared to listeners who were exposed to a narrow distribution.

In a later study D. F. Kleinschmidt and Jaeger (2016) tested listener response to talker

statistics by shifting the means of the voiced and voiceless categories between conditions. Specifically, the mean values for /b/ and /p/ were shifted rightwards by several magnitudes, as well as leftwards, from the expected mean values of a typical American English talker while the category variances remained identical and the distance between the category means were kept constant. With this manipulation of means they were able to investigate how inclined listeners are to adapt their categorisation behaviors when the statistics of the exposure talker were shifted beyond the bounds of a typical talker.

In all exposure conditions, listeners on average adapted to the exposure talker by shifting their categorization function in the direction of the predicted function of an ideal listener (a listener who perfectly learned the exposure talker's cue statistics). However, in all conditions, listener categorization fell short of the predicted ideal categorization boundary. This difference between the observed and predicted categorization functions was larger, the greater the magnitude of the shift from the typical talker's distribution, suggesting some constraints on adaptation.

The study we report here builds on the pioneering work of Clayards et al. (2008) and D. F. Kleinschmidt and Jaeger (2016) with the aim to shed more light on the role of prior implicit knowledge on adaptation to an unfamiliar talker.

Specifically, while K&J16 demonstrated how prior beliefs of listeners can be inferred computationally from post-exposure categorisation, their experiment was not designed to capture listener categorisation data before exposure to a novel talker. Nor did they run intermittent tests to scrutinise the progress of adaptation. In the ideal adapter framework, listener expectations are predicted to be rationally updated through integration with the incoming speech input and thus can theoretically be analysed on a trial-by-trial basis. The overall design of the studies reported here were motivated by our aim to understand this incremental belief-updating process which has not been closely studied in previous work. We thus address the limitations of previous work and in conjunction, make use of ideal observer models to validate baseline assumptions that accompany this kind of speech perception study – that listeners hold prior expectations or beliefs about cue distributions based on previously experienced speech input (here taken to mean native AE listeners' lifetime of experience with AE). Arriving at a definitive conclusion of what shape and form those beliefs take is beyond the scope of this study however we attempt to explore the

various proposals that have emerged from more than half a century of speech perception research.

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., 2008) 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).

2 Experiment 1: Listener's expectations prior to informative exposure

Experiment 1 investigates native (L1) US English listeners' categorization of word-initial stop voicing by an unfamiliar female L1 US English talker, prior to more informative exposure. Specifically, listeners heard isolated recordings from a /d/-/t/ continuum, and had to respond which word they heard (e.g., "din" or "tin"). The recordings varied in voice onset time (VOT), the primary phonetic cue to word-initial stop voicing in L1 US English, as well as correlated secondary cues (f0 and rhyme duration). Critically, exposure was relatively uninformative about the talker's use of the phonetic cues in that all phonetic realizations occurred equally often. The design of Experiment 1 serves two goals.

The first goal is methodological. We use Experiment 1 to test basic assumptions about the paradigm and stimuli we employ in the remainder of this study. Experiment 1 provides perceptual norms for a new set of /d/-/t/ stimuli developed to improve ecological validity. While it is well-established that larger VOT values make it more likely that listeners categorize a

recording as having the voiceless stop (here "t"), the specific categorization function can vary between talkers (**REFS?**) and listeners (e.g., Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016). For Experiment 2, we wanted to have an estimate of the category boundary between /d/ and /t/, as perceived by listeners *of the type we seek to recruit for Experiment 2* and specifically *for the stimulus recordings we employ in all experiments*. Categorization data from this first experiment would also reveal differences in how listeners perceived the four minimal pairs if any. Although we had no reason to expect them to elicit substantially different behaviour in participants we did intend to use only three of the four pairs available, in experiment 2 in order to be consistent with the previous studies that we wish to extend. In relation to establishing norms, we also aimed to test whether listeners' categorization behavior changes over time even when exposure is relatively uninformative—i.e., even when the stimuli listeners hear form a uniform distribution across trials. If listeners exhibit changes in categorization behavior even for such uninformative input, this would have implications for the interpretation of adaptive behavior when input like the type we employ in Experiment 2 is actually informative .

The second purpose of Experiment 1 is to introduce and illustrate relevant theory. We compare different models of listeners' prior expectations against listeners' categorization responses in Experiment 1. The different models all aim to capture the implicit expectations of an L1 adult listener of US English might have about the mapping from acoustic cues to /d/ and /t/ based on previously experienced speech input. As we describe in more detail after the presentation of the experiment, the models differ, however, in whether these prior expectations take into account that talkers can differ in the way they realize /d/ and /t/. This ability to take into account talker differences even prior to more informative exposure is predicted—though through qualitatively different mechanisms, as we discuss below—both by normalization accounts (Cole, Linebaugh, Munson, & McMurray, 2010; McMurray & Jongman, 2011) and by accounts that attribute adaptive speech perception to changes in category representations (Bayesian ideal adaptor theory, D. F. Kleinschmidt & Jaeger, 2015; EARSHOT, Magnuson et al., 2020; episodic theory, Goldinger, 1998; exemplar theory, Johnson, 1997; Pierrehumbert, 2001). It is, however, unexpected under accounts that attribute adaptive speech perception solely to ad-hoc changes in decision-making. We did not expect Experiment 1 to yield a decisive conclusion with regard to

this second goal, which is also addressed in Experiment 2. Rather, we use Experiment 1 as a presentationally convenient way to introduce some of the different models and provide readers with initial intuitions about what experiments of this type can and cannot achieve.

2.1 Methods

2.1.1 Participants

Participants were recruited over Amazon's Mechanical Turk platform, and paid $2.50 each (for a targeted remuneration of $6/hour). The experiment was only visible to Mechanical Turk participants who (1) had an IP address in the United States, (2) had an approval rating of 95% based on at least 50 previous assignments, and (3) had not previously participated in any experiment on stop voicing from our lab.

24 L1 US English listeners (female = 9; mean age = 36.2 years; SD age = 9.2 years) completed the experiment. To be eligible, participants had to confirm that they (1) spent at least the first 10 years of their life in the US speaking only English, (2) were in a quiet place, and (3) wore in-ear or over-the-ears headphones that cost at least $15.

2.1.2 Materials

We recorded multiple tokens of four minimal word pairs ("dill"/"till", "dim"/"tim", "din"/"tin", and "dip"/"tip") from a 23-year-old, female L1 US English talker with a mid-Western accent. These recordings were used to create four natural-sounding minimal pair VOT continua (dill-till, dip-tip, din-tin, and dip-tip) using a Praat script (Winn, 2020). The full procedure is described in the supplementary information (SI, **??**). The VOT continua ranged from -100ms VOT to +130ms VOT in 5ms steps. Experiment 1 employs 24 of these steps (-100, -50, -10, 5 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 100, 110, 120, 130).

We further set the F0 at vowel onset to follow the speaker's natural correlation which was estimated through a linear regression analysis of all the recorded speech tokens. We did this so that we could determine the approximate corresponding f0 values at each VOT value along the continua as predicted by this talker's VOT. The duration of the vowel was set to follow the natural trade-off relation with VOT reported in Allen and Miller (1999). This approach closely resembles

that taken in Theodore and Monto (2019), and resulted in continuum steps that sound highly natural (unlike the robotic-sounding stimuli employed in Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016). All stimuli are available as part of the OSF repository for this article.

In addition to the 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 as catch trials. Stimuli intensity were standardised at 70 dB sound pressure level.

2.1.3 Procedure

Participants underwent a headphones test and gave their consent to the study following the guidelines of the Research Subjects Review Board of the University of Rochester. After participants passed the headphones test and gave their consent they were taken to an instructions page. They were informed that they would hear a female talker say a word and that they would have to select which word they heard. They were instructed 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. This was done to encourage their full attention to the playbacks.

Each trial started with a green fixation dot being displayed. At 500ms from trial onset, two minimal pair words appeared on the screen, as shown in Figure 1. At 1000ms from trial onset, an audio recording from the matching minimal pair continuum started playing. Participants were required to click on the word they heard. After participants clicked on the word, the next trial began.

The trials were presented randomly and the order of the written word forms were counter-balanced across participants. Participants were given 192 target trials along a 24-step VOT continuum to categorise. VOT tokens in the lower and upper ends were distributed over larger increments because stimuli in those ranges were expected to elicit floor and ceiling effects, respectively. Each word pair was played twice at each VOT step, constituting a uniform distribution. In addition to the critical trials, 12 catch trials were inserted randomly throughout the experiment. These trials served as a check on participant attention throughout the experiment. Participants were given the opportunity to take breaks after every 60 trials.

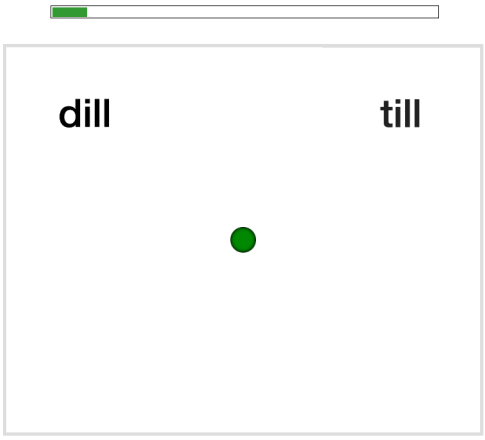

Figure 1. Example trial display. The words were displayed 500ms after trial onset and the audio recording of the word was played 1000ms after trial onset

Participants spent an average of 12 minutes (SD = 4.8) on the trials after which they answered a short survey about the experiment.

2.2 Results

We first present the behavioral analyses of listeners' categorisation responses, and then compare them to predictions of ideal observers.

2.2.1 Exclusions

We excluded from analysis participants who committed more than 3 errors out of the 12 catch trials (<75% accuracy, N = 3), participants with an average reaction time (RT) more than three standard deviations from the mean of the by-participant means (N = 0), and participants who reported not to have used headphones (N = 0) or not to be native (L1) speakers of US English (N = 0). For the remaining participants, trials that were more than three SDs from the participant's mean RT were excluded from analysis (1.6%). Finally, we excluded participants (N = 0) who had less than 50% data remaining after these exclusions.

2.2.2 Behavioral analyses

The goal of our behavioral analyses was to address three methodological questions that are of relevance to Experiment 2: (1) whether our stimuli resulted in 'reasonable' categorisation

functions, (2) whether these functions differed between the four minimal pair items, and (3) whether participants' categorisation functions changed throughout the 192 test trials.

To address these questions, we fit a single Bayesian mixed-effects psychometric model to participants' categorization responses on critical trials (e.g., **prins2011?**). The model describes the probability of "t"-responses as a weighted mixture of a lapsing-model and a perceptual model. The lapsing model is a mixed-effects logistic regression (Jaeger, 2008) that predicts participant responses that are made independent of the stimulus—for example, responses that result from attentional lapses. These responses are independent of the stimulus, and depend only on participants' response bias. The perceptual model is a mixed-effects logistic regression that predicts all other responses, and captures stimulus-dependent aspects of participants' responses. The relative weight of the two models is determined by the lapse rate, which is described by a third mixed-effects logistic regression.

The *lapsing model* only had an intercept (the response bias in log-odds) and by-participant random intercepts. Similarly, the *model for the lapse rate* only had an intercept (the lapse rate) and by-participants random intercepts. Previous studies with similar paradigms have typically found lapse rates of 0-10% ($<$ -2.2 log-odds, e.g., Clayards et al., 2008; D. F. Kleinschmidt & Jaeger, 2016). No by-item random effects were included for the lapse rate nor lapsing model since these parts of the analysis—by definition—describe stimulus-*in*dependent behavior. The *perceptual model* included an intercept and VOT, as well as the full random effect structure by participants and items (the four minimal pair continua), including random intercepts and random slopes by participant and minimal pair. We did not model the random effects of trial to reduce model complexity. This however makes our analysis of trials in the model anti-conservative.

Based on previous experiments, we expected a strong positive effect of VOT, with increasing proportions of "t"-responses for increasing VOTs. We did not have clear expectations for the effect of trial other than that responses should become more uniformed (i.e move towards 50-50 "d"/"t"-bias or 0-log-odds) as the experiment progressed (Liu & Jaeger, 2018a) due to the un-informativeness of the stimuli. Finally, the models included the covariance between by-participant random effects across the three linear predictors for the lapsing model, lapse rate model, and perceptual model. This allows us to capture whether participants who lapse more

often have, for example, different response biases or different sensitivity to VOT (after accounting for lapsing).

We fit the model using the package `brms` (Bürkner, 2017) in R (R Core Team, 2021a; RStudio Team, 2020). Following our previous work (Hörberg & Jaeger, 2021; X. Xie et al., 2021), we used weakly regularizing priors to facilitate model convergence. For fixed effect parameters, we standardized continuous predictors (VOT) by dividing through twice their standard deviation (**gelman2008standardize?**), and used Student priors centered around zero with a scale of 2.5 units (following **gelman2008weakly?**) 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 $\hat{R}$ were close to 1.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

The lapse rate was estimated to be on the slightly larger side, but within the expected range (7.5 %, 95%-CI: 2.3 to 20.4%; Bayes factor: Inf 90%-CI : -3.49 to -1.55). Maximum a posteriori (MAP) estimates of by-participant lapse rates ranged from XX. Very high lapse rates were estimated for four of the participants with one in particular whose CI indicated exceptionally high uncertainty. These lapse rates might reflect data quality issues with Mechanical Turk that started to emerge over recent years (see **REFS?**; and, specifically for experiments on speech perception, **cummings2023?**), and we return to this issue in Experiment 2.

The response bias were estimated to slightly favor “t”-responses (54.8 %, 95%-CI: 17.7 to 82.5%; Bayes factor: 1.69 90%-CI : -1.17 to 1.33), as also visible in Figure 2 (left). Unsurprisingly, the psychometric model suggests high uncertainty about the participant-specific response biases, as it is difficult to reliably estimate participant-specific biases while also accounting for trial and VOT effects (range of by-participant MAP estimates: XX). For all but four participants, the 95% CI includes the hypothesis that responses were unbiased. Of the remaining four participants, three were biased towards “t”-responses and one was biased toward “d”-responses.

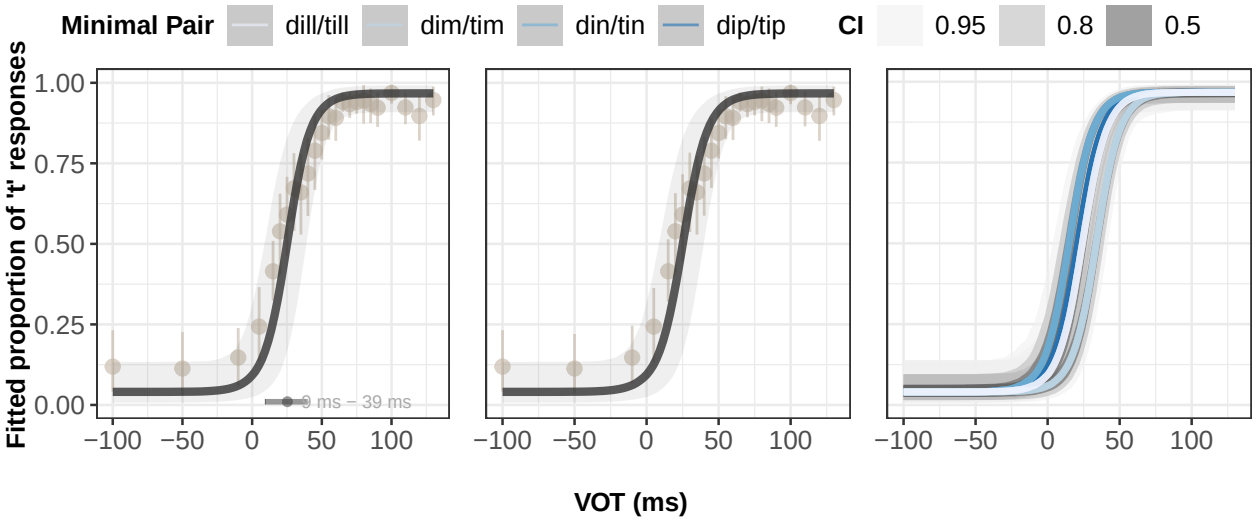

Figure 2. Fitted categorisation functions to listeners' responses marginalizing over trial effects as well as all random effects (left) and combined effects of VOT and trial (middle), marginalizing over all random effects. Vertical point ranges represent the mean proportion of t-responses at respective VOTs and vertical bars denote the 95% bootstrapped confidence interval. Black error bar (bottom of left panel) denotes the 95% quantile interval of the points of subjective equality (PSE), derived from 8000 sample draws from the posterior distribution of estimated population parameters. Rightmost plot shows the predicted categorisation functionsfor all four minimal pair items.

There was no convincing evidence of a main effect of trial ($\hat{\beta}$ = -0.2 95%-CI: -0.7 to 0.4; Bayes factor: 2.67 90%-CI : -0.58 to 0.27). Given the slight overall bias towards "t"-responses, the direction of this effect indicates that participants converged towards a 50/50 bias as the test phase proceeded. This is also evident in Figure 2 (right). In contrast, there was clear evidence for a positive main effect of VOT on the proportion of "t"-responses ($\hat{\beta}$ = 12.6 95%-CI: 9.8 to 15.6; Bayes factor: Inf 90%-CI : 10.29 to 15.03). The effect of VOT was consistent across all minimal pair words as evident from the slopes of the fitted lines by minimal pair 2 (left). MAP estimates of by minimal pair slopes ranged from . The by minimal-pair intercepts were more varied (MAP estimates:) with one of the pairs, dim/tim having a slightly lower intercept resulting in fewer 't'-responses on average. In all, this justifies our assumptions that word pair would not have a substantial effect on categorisation behaviour. From the parameter estimates of the overall fit we obtained the category boundary from the point of subjective equality (PSE) (25ms) which we use for the design of Experiment 2.

Finally to accomplish the first goal of experiment 1, we look at the interaction between

VOT and trial. There was weak evidence that the effect of VOT decreased across trials ($\hat{\beta}$ = -0.6 95%-CI: -2.6 to 1.5; Bayes factor: 2.56 90%-CI : -2.3 to 1.12). The direction of this change—towards more shallow VOT slopes as the experiment progressed—makes sense since the test stimuli were not informative about the talker's pronunciation. Similar changes throughout prolonged testing have been reported in previous work. (Liu & Jaeger, 2018b, 2019; **REFS?**).

Overall, there was little evidence that participants substantially changed their categorisation behaviour as the experiment progressed. Still, to err on the cautious side, Experiment 2 employs shorter test phases.

2.3 Comparisons to model of adaptive speech perception

We now turn to final aim of experiment 1 which is to make use of computational models to delve into the theoretical underpinnings that inform the assumptions we make in studies of this kind.

Speakers' productions can act as a proxy for listeners' implicit knowledge of the distributional patterns of cues. This production-perception relationship within a phonological system was observed in early work by (Abramson & Lisker, 1973) who found that production statistics of talkers along VOT aligned well with data from listeners who had categorised a separate set of synthesised VOT stimuli. This allows for the use of analytic models as tools for predicting categorisation behaviour from speech production (Nearey & Hogan, 1986).

We apply this principle in fitting ideal observer (IO) models by linking the distributional patterns of input to the categorisation behaviour that listeners make in the perception of our stimuli. We compare the categorisation behaviour against predictions of several IO models differentiated by the various assumptions they incorporate. These IOs are trained on cue measurements extracted from an annotated database of 92 L1 US-English talkers' productions (Chodroff & Wilson, 2017) of word initial stops. By using IOs trained solely on production data to predict behaviour we avoid additional computational degrees of freedom and limit the risk of overfitting the model to the data thus reducing bias.

Hypotheses about the nature of long-term representations maintained by listeners continues to be debated and revised. On one hand there is the proposition that automatic processes that operate purely on the acoustic input is sufficient mechanism for listeners to cope with variation;