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 affect 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 Theodore & Miller, 2010) 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, 2008a; D. F. Kleinschmidt & Jaeger, 2016; and Theodore & Monto, 2019).

Clayards et al. (2008a) 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. (2008a) 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., 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).

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. We obtain estimates of the category boundary between /d/ and /t/ *for the specific stimuli used in Experiment 2*, as perceived by *the type of listeners we seek to recruit for Experiment 2*. We also test whether prolonged

testing across the phonetic continuum changes listeners' categorization behavior. Previous work has found that prolonged testing on uniform distributions can reduce the effects of previous exposure (Liu & Jaeger, 2018a; e.g., **mitterer2011?**), at least in listeners of the age group we recruit from (**scharenborg-janse2013?**). However, these studies employed only a small number of 5-7 perceptually highly ambiguous stimuli, each repeated many times. In Experiment 1, we employ a much larger set of stimuli that span the entire continuum from very clear /d/s to very clear /t/s, each presented only twice. If prolonged testing changes listeners' responses, this has to be taken into account in the design of Experiment 2.

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 that Experiment 1 yields 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). 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.

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., 2008a; D. F. Kleinschmidt & Jaeger, 2016). All stimuli are available as part of the OSF repository for this

article.

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 as catch trials. Stimulus intensity was set to 70 dB sound pressure level for all recordings.

2.1.3 Procedure

The code for the experiment is available as part of the OSF repository for this article. A live version is available at (https://www.hlp.rochester.edu/FILLIN-FULL-URL). The first page of the experiment informed participants of their rights and the requirements for the experiment: that they had to be native listeners of English, wear headphones for the entire duration of the experiment, and be in a quiet room without distractions. Participants had to pass a headphone test, and were asked to keep the volume unchanged throughout the experiment. Participants could only advance to the start of the experiment by acknowledging each requirement and consenting to the guidelines of the Research Subjects Review Board of the University of Rochester.

On the next page, participants were informed about the task for the remainder of the experiment. They were informed that they would heard a female talker speak a single word on each trial, and had to select which word they heard. Participants 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.

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. For each participant, /d/-initial words were either always displayed on the left side or always displayed on the right side. Across participants, this ordering was counter-balanced. After participants clicked on the word, the next trial began.

Participants heard 192 target trials (four minimal pair continua, each with 24 VOT steps, each heard twice). In addition, participants heard 12 catch trials. On catch trials, participant saw

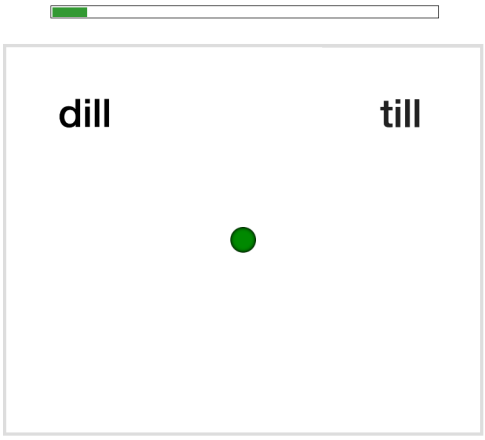

Figure 1. Example trial display. The words were displayed 500ms after trial onset and the audio recording of the word was played 1000ms after trial onset

two written catch stimuli on the screen (e.g., "flare" and "rare"), and heard one of them (e.g. "rare"). Since these recordings were easily distinguishable, they served as a check on participant attention throughout the experiment.

The order of trials was randomized for each participant with the only constraint that no stimulus was repeated before each stimulus had been heard at least once. Catch trials were distributed randomly throughout the experiment with the constraint that no more than two catch trials would occur in a row. Participants were given the opportunity to take breaks after every 60 trials. Participants took an average of 12 minutes (SD = 4.8) to complete the 204 trials, after which they answered a short survey about the experiment.

2.1.4 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 Behavioral results

We first present the behavioral analyses of participants' categorisation responses. Then we compare participants' responses to the predictions of different models fit on the distribution of stop voicing cues in a large database of L1 US English productions of word-initial /d/s and /t/s (Chodroff & Wilson, 2018).

2.2.1 Analysis approach

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?**). This model is essentially an extension of mixed-effects logistic regression that also takes into account attentional lapses. A failure to do so—while commonplace in research on speech perception (incl. our own work, but see Clayards, Tanenhaus, Aslin, & Jacobs, 2008b; D. F. Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (e.g., **wichman-hill2001?**). The mixed-effects psychometric 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 contained 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. 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 potentially makes our analysis of trials in the model anti-conservative. 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 previous work from our lab (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.

2.2.2 Expectations

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, 2018b) due to the un-informativeness of the stimuli. Previous studies with similar paradigms have typically found lapse rates of 0-10% (< -2.2 log-odds, e.g., Clayards et al., 2008a; D. F. Kleinschmidt & Jaeger, 2016).