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; 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 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 acoustic cue-to- phonetic 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 aligns with the statistics of exposure stimuli Theodore & Monto (2019) suggesting a possible change in cue-to-category mappings.

In Clayards et al. (2008a) listeners responded with greater uncertainty after they were exposed VOT distributions of a "beach-peach" contrast that had wider variances relative to another group who had heard the same contrasts distributed over a narrower variance. 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 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 in varying durations, 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.

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 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 distribuions of cues which they bring to an encounter. The strength of these beliefs has bearing on their propensity to adapt to a new talker. Listeners' strengths in prior expectations are represented by parameters in the model. The behaviour observed collectively in all experiments so far should be able to indicate roughly what the parameter values are. It has been shown in Kleinschmidt and Jaeger (2016) that adaptation is constrained – does this i

–WHAT'S NEW HERE– The study we report here builds on the pioneering work of Clayards et al. (2008a) and 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 (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).

1.1 Methods

1.1.1 Participants

Participants were recruited over the Prolific platform and experiment data (but not participant profile data) were collected, stored, and via proliferate ((**schuster?**)). They were paid $8.00 each (for a targeted remuneration of $9.60/hour). The experiment was visible to participants following a selection of Prolific's available pre-screening criteria. Participants had to (1) have US nationality, (2) report to only know English, and (3) had not previously participated in any experiment from our lab on Prolific.

126 L1 US English listeners (male = 60, female = 59, NA = 3; mean age = 38 years; SD age = 12 years) completed the experiment. Due to data transfer errors 4 participants' data were not stored and therefore not included in this analysis. 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 free from distractions, and (3) wore in-ear or over-the-ears headphones that cost at least $15.

1.1.2 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 from New Hampshire, judged to have a "general American" 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). 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. The full procedure is described in the supplementary information (SI, **??**).

We also 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 resulted in continuum steps that sound highly natural (unlike the robotic-sounding stimuli employed in Clayards et al., 2008a; Kleinschmidt & Jaeger, 2016). All stimuli are available as part of the OSF repository for this article.

Prior to creating the three exposure conditions of the experiment, we ran a norming experiment to test US-L1 listeners' perception of our stimuli and to determine a baseline categorisation boundary for this talker. The norming experiment also served as a measure to detect possible anomalous features present in our stimuli (for e.g. if it would elicit unusual categorisation behaviour or whether certain minimal-pairs had an exaggerated effect on categorisation). For the norming experiment the VOT continua employed 24 VOT steps ranging from -100ms VOT to +130ms (-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 found VOT to have the expected effect on the proportion of "t"-responses, i.e. higher VOTs elicited greater "t"-responses and that the word-pairs did not differ substantially from each other. The results and analysis of the norming experiment are reported in full in section **??**.

A subset of the materials were used to generate the three exposure conditions; in particular three continua of the minimal pairs, dill-till, din-tin, and dip-tip. The dim-tim continuum was omitted in order to keep the pairs as distinct as possible.

We employed a multi-block exposure-test design **??** which enabled the assessment of listener perception before informative exposure as well as incrementally at intervals during informative exposure (after every 48 exposure trials). To have a comparable test between blocks and across conditions, test blocks were made up of a uniform distribution of 12 VOT stimuli (-5, 5, 15, 25, 30, 35, 40, 45, 50, 55, 65, 70), identical across test blocks and between conditions. Each of the test tokens were presented once at random. The test blocks were kept short to minimise distortion of the intended distribution to be presented by the end of the exposure phase. After the final exposure block we tripled the number of test blocks to increase the statistical power to detect exposure induced behavioural changes.

The conditions were created by first generating the baseline distribution (+0ms shift) and then shifting that distribution by +10ms and by +40ms to the right of the VOT continuum to create the remaining two conditions.

To construct the +0ms shift exposure distribution we first computed the point of subjective equality (PSE) from the perceptual component of the fitted psychometric function of listener responses in the norming experiment. The PSE corresponds to the VOT duration that was perceived as most ambiguous across all participants during norming (i.e. the stimulus that on average, elicited equal chance of being categorised as /d/ or /t/) thus marking the categorical boundary. From a distributional perspective the PSE is where the likelihoods of both categories intersect and have equal probability density (we assumed Gaussian distributions and equal prior probability for each category) [SOMETHING HERE ABOUT GAUSSIANS BEING A CONVENIENT ASSUMPTION?]. To limit the infinite combinations of category likelihoods that could intersect at this value, we set the variances of the /d/ (80ms) and /t/ (270ms (lowered from 398 because of dip-tip pair limitations)) categories based on parameter estimates (X. Xie, Jaeger, and Kurumada (2022)) obtained from the production database of word-initial stops in Chodroff and Wilson (2017). To each variance value we added 80ms following (Kronrod, Coppess, and Feldman (2016)) to account for variability due to perceptual noise since these likelihoods were estimated from perceptual data. We took an additional degree of freedom of setting the *distance between the means* of the categories at 46ms; this too was based on the mean for /d/ and /t/ estimated from the production database. The means of both categories were then obtained through a grid-search process to find the likelihood distributions that crossed at 25ms VOT (see XX of SI for further detail on this procedure).

The distributional make up was determined through a process of sampling tokens from a discretised normal distribution with values rounded to the nearest multiple of 5 integer (available through the `extraDistr` package in R). For each exposure block 8 VOT tokens per minimal word pair were sampled from discrete normal distributions of each category of the +0ms condition, giving 24 /d/ and 24 /t/ items (48 critical trials) per block. Additionally, each exposure block contained 2 instances of 3 catch items, giving 6 catch trials per block. The sampled distributions of VOT tokens were increased by a margin of +10ms and +40 ms to create the remaining two conditions. Three variants of each condition list were created so that exposure blocks followed a latin-square order.

Lastly, half of the exposure trials were randomly assigned as labelled trials. In labelled trials, participants receive clear information of the word's category as both orthographic options will always begin with the intended sound. For example if a trial was intended to be "dill" then the two image options will either be "dill" and "dip" or "dill" and "din". Test trials were always *unlabelled*.

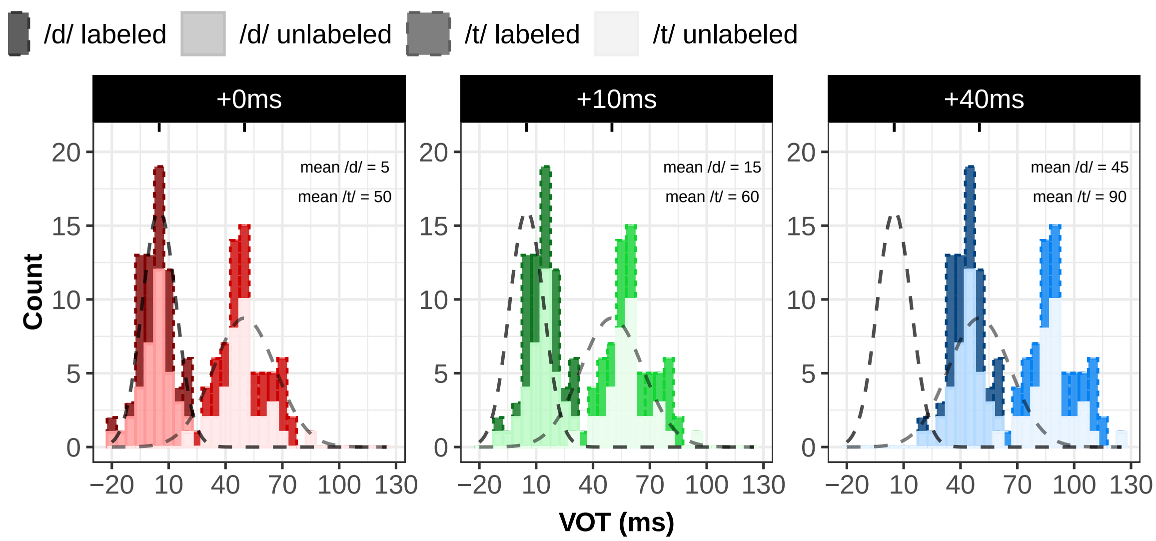

Figure 1

1.1.3 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 hear a female talker speak a single word on each trial, and had to select which word they heard. They were also informed that they needed to click a green button that would be displayed during each trial when it “lights up” in order to hear the recording of the speaker saying the word. 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 dark-shaded green fixation dot being displayed. At 500ms from

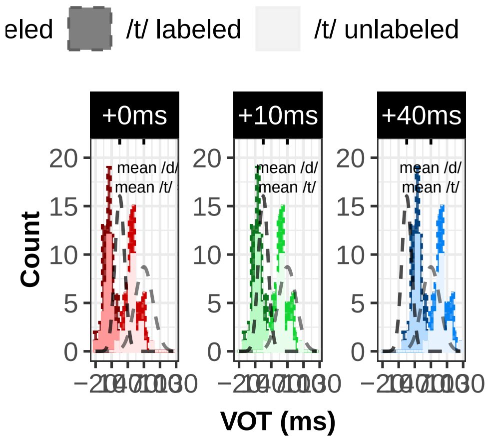

Figure 2

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. Participants responded by clicking on the word they heard and the next trial would begin. The placement of the word presentations were counter-balanced across participants.

Participants underwent 234 trials which included 6 catch trials in each exposure block (18 in total). Since these recordings were easily distinguishable, they served as a check on participant attention throughout the experiment. 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 during exposure blocks. Participants took an average of 17 minutes (SD = 9) to complete the 234 trials, after which they answered a short survey about the experiment.

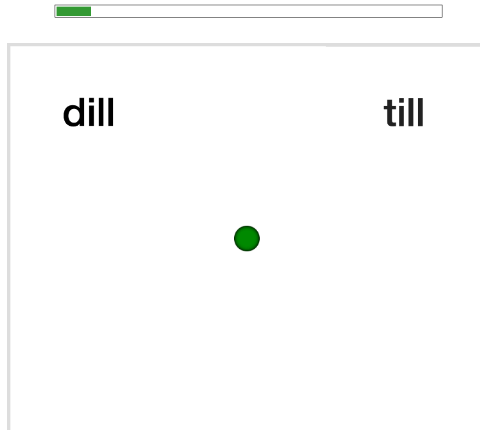

Figure 3. Example trial display. The words were displayed 500ms after trial onset. The green button would turn bright green signalling participants to click on the dot to play the recording.

1.1.4 Exclusions

We excluded from analysis participants who committed more than 3 errors out of the 18 catch trials (<84% 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 (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).

In addition, participants' categorization during the early phase of the experiment were scrutinised for their slope orientation and their proportion of "t"-responses at the least ambiguous locations of the VOT continuum. The early phase of the experiment was defined as the first 36 trials and the least ambiguous locations were defined as -20ms below the empirical mean of the /d/ category and +20ms above the empirical mean of the /t/ category. These means were obtained from the production data estimates by X. Xie et al. (2022).

1.1.5 Analysis approach

##Results ## Regression analysis The regression analysis addresses two main questions: Do participants shift their categorisation behaviour in an incremental fashion, i.e. do they exhibit categorisation behaviour that draws closer to the ideal categorisation function with each successive exposure block? Are the differences in shifts between the conditions proportional to the magnitude of the shifts between exposure distributions i.e. is the PSE of the +40ms condition 3 times that of the +10ms condition?

We fit a Bayesian mixed-effects psychometric model with lapse and perceptual components. Continuous predictors were standardised to twice the standard deviation and priors and sampling parameters were identical to those specified in experiment 1.

To analyse the incremental effects of exposure condition on the proportion of /t/ responses at test, the perceptual model contained 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 the sixth test block), VOT (scaled to twice the), and their full factorial interaction. For the perceptual model, “t”-responses were regressed on the three-way interaction of VOT, condition, and block. Random effects were modelled with varying intercepts and slopes by participant and varying intercepts and slopes by minimal pair item. The lapsing model which estimates participant bias on trials with attention lapses was fitted without an intercept but with an offset [how does one describe this? what does offset(0) represent]. Finally, a population-level intercept was fitted to estimate the lapse rate. Random effects for the lapsing model and lapse rates were not fitted to limit the number of parameters and to ensure model convergence.

1.1.6 Expectations

Given previous findings of Kleinschmidt and Jaeger (2016) we expected participants in the various exposure conditions to shift their average categorization functions towards the direction of the ideal categorization function implied by their respective exposure distributions. We expected the differences between the groups to be most pronounced after the final exposure block as they would have had the complete exposure to all the tokens that make up the exposure distributions. This follows from predictions of incremental Bayesian belief-updating – that listeners would integrate their prior expectations with the current input to infer the present talker’s cue-to-category-mapping (the posterior distribution). Also based on previous findings, we expected the +40ms group to not fully converge on the ideal categorization function as it was previously found that the further an exposure talker's cue distributions deviated from a *typical* talker's, the further the distance of categorization function from the ideal boundary. We therefore expected to see differences in categorizations between the +10ms and +40ms conditions such that listeners in the +40ms condition would shift more than those in the +10ms condition but to have an average categorization function located to the left of the ideal function. (Kleinschmidt & Jaeger, 2016).

1.2 Behavioral results

1.2.1 Analysis approach

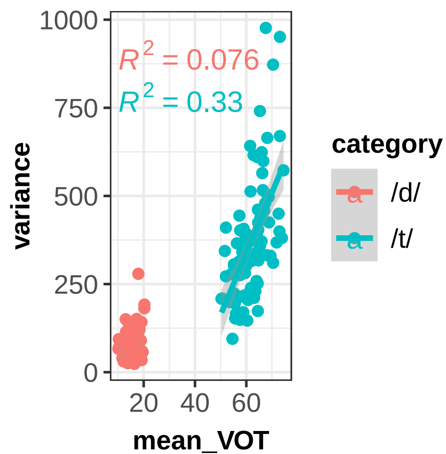

Figure 4

1.3 Regression analysis

The regression analysis addresses two main questions: Do participants shift their categorisation behaviour in an incremental fashion, such that the categorisation function draws closer to the ideal categorisation function with each successive exposure block? Are the differences in shifts between the conditions proportional to the magnitude of the shifts between exposure distributions i.e. is the PSE of the +40ms condition 3 times that of the +10ms condition?

1.3.1 Expectations

Given previous findings of Kleinschmidt and Jaeger (2016) we expected participants in the various exposure conditions to shift their average categorization functions towards the direction of