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

Maryann Tan[1, 2] & T. Florian Jaeger[2,3]

[1] Centre for Research on Bilingualism, University of Stockholm

[2] Brain and Cognitive Sciences, University of Rochester

[3] Computer Science, University of Rochester

Author Note

We are grateful to ### ommitted for review ###

Correspondence concerning this article should be addressed to Maryann Tan, Department of Bilingualism, Stockholm University, Sweden. E-mail: maryann.tan@biling.su.se

Abstract

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

Keywords: speech perception; adaptation; incremental changes; distributional learning

Word count: X

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

1 TO-DO

1.1 Highest priority

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

        you deal with formatting the conditions and various names floating around (Shift0, 10, etc.; +0, +10, etc.; baseline, + 10 etc.). Pick one, do it at the top of the pipeline (i.e., in the input script). This will reduce the potential for error in your own coding, make your code in the main paper shorter, and it'll be much easier to read for others trying to follow your code (including me).

        * Remove all data formatting code from the paper Rmd. There should only be a single load line.
        * I've moved the code loading the chodroff data into the new pre-amble.R file. Consider doing the same for the experiment data. That way the data that we need throughout are available throughout.

- Clean up functions.R file:

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

- Try to set local constants at top of chunk. e.g., Don't have stuff like empirical_means <- c(17, 62) in the middle of a chunk.
- It's best not to save unnecessary objects but if you do, remove them after they are no longer needed (e.g., the various excl.headphone, etc. in section 2: you could just have that code inline without ever storing them. But it's ok to do things the way you do. Just remove them after they have done their job).

1.2 Medium priority

- MARYANN
- FLORIAN
- think about table 1 and 2: how to change the wording on tables to actually refer to intercepts rather than PSEs or change the figures? Changing current representations of analyses to improve intuitive-ity.
- write overview of results
- restructure results presentation.
- write SI sections with proofs

1.2.1 Lower Priority

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

1.3 To do later

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

integrate: magnuson-nusbaum2007; magnuson2020

1 Introduction

Adaptivity is a hallmark of human speech perception. When exposed to an unfamiliar accent, listeners might initially experience processing difficulty, but this difficulty alleviates with exposure, supporting faster and more accurate speech recognition (e.g., Bradlow & Bent, 2008; **bradlow2023?**; **clarke-garrett2004?**; **sidaras2009?**; **xie2018jasa?**; for review, see **baeseberk2018?**; **xie2021jep?**). Such adaptivity is now recognized as a critical component of robust language understanding. Research over the last few decades has made strides in identifying its fundamental properties, including the conditions required for successful adaptation, its generalizability across talkers, and its longevity (Cummings & Theodore, 2023; for reviews, see **bent-baeseberk2021?**). Progress has also been made in understanding how adaptation proceeds. In particular, it is now clear that listeners' categorization function—the mapping from acoustic or phonetic cues to phonetic categories—changes based on the phonetic properties of recent input (Bertelson, Vroomen, & De Gelder, 2003; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Idemaru & Holt, 2011; McMurray & Jongman, 2011; Norris, McQueen, & Cutler, 2003; Reinisch & Holt, 2014; **cole2011?**; **eisner-mcqueen2005?**; **kraljic-samuel2005?**; **kurumada2013?**; **xie2018jep?**; for review, see Schertz & Clare, 2020; Xie, Jaeger, & Kurumada, 2023). This has led to the development of stronger theories and models of adaptive speech perception that explicitly link the phonetic properties of recent speech input to changes in subsequent speech recognition (e.g., Apfelbaum & McMurray, 2015; Harmon, Idemaru, & Kapatsinski, 2019; Johnson, 1997; Kleinschmidt & Jaeger, 2015; **nearey-assman2007?**; **winter-lancia2013?**; **sohoglu-davis2016?**).

While these theories differ in important aspects (for review, see Xie et al., 2023), they share critical predictions that have remained largely untested. In particular, adaptation is predicted to develop *incrementally*, *accumulate* over exposure, at each point depending *gradiently* on both the amount and the statistics of the speech input experienced so far. We report initial results from a novel repeated exposure-test paradigm that aims to test these predictions during the early moments of adaptation. In the long run, a clearer picture of how speech recognition changes with exposure, how those changes accumulate, and whether there are constraints on this accumulation has been identified as critical in developing stronger tests for theories of adaptive speech perception (Xie et al., 2023; **martin2023?**).

The experiment we report builds on computational and behavioral findings in research on unsupervised learning during speech perception (Clayards et al., 2008; Kleinschmidt & Jaeger, 2016) and visually- or lexically-guided perceptual learning (Cummings & Theodore, 2023; Vroomen, Linden, De Gelder, & Bertelson, 2007; **kleinschmidt-jaeger2012?**). The two paradigms have complementing strengths. Perhaps the clearest evidence that adaptation to unfamiliar speech depends on the statistics of the input—specifically, the distribution of phonetic cues—comes from the former paradigm (Bejjanki, Beck, Lu, & Pouget, 2011; Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016; Munson, 2011; Nixon, Rij, Mok, Baayen, & Chen, 2016; Theodore & Monto, 2019). In an important early study, Clayards and colleagues exposed two different groups of US English listeners to instances of "b" and "p" that differed in their distribution along the voice onset time continuum (VOT). VOT is the primary phonetic cue to word-initial /b/-/p/, /d/-/t/, /g/-/k/ in US English: the voiced category (e.g. /b/) is produced with lower VOT than the voiceless category (e.g., /p/). Clayards and colleagues held the VOT means of /b/ and /p/ constant between the two exposure groups, but manipulated whether both /b/ and /p/ had wide or narrow variance along VOT. Exposure was unlabeled: on any trial, listeners saw pictures of, e.g., bees and peas on the screen while hearing a synthesized recording along the "bees"-"peas" continuum (obtained by manipulating VOT). Listeners' task was to click on the picture corresponding to the word they heard. If listeners adapt by learning the category statistics of the exposure input—in this case, the distribution of VOT for /b/ and /p/—they were predicted to change their categorization function along VOT such that listeners in the wide variance group should exhibit a more shallow categorization function than the narrow variance group. This is precisely what Clayards and colleagues found (see also Nixon et al., 2016; Theodore & Monto, 2019). Together with more recent findings from adaptation to natural accents (Hitczenko & Feldman, 2016; Tan, Xie, & Jaeger, 2021; **xie2021cognition?**), this suggests that the *outcome* of adaptation qualitatively follows the predictions of distributional learning models (e.g., exemplar theory, Johnson, 1997; ideal adaptors, Kleinschmidt & Jaeger, 2015).

It leaves open, however, how adaptation *incrementally accumulates* with increasing exposure, and whether it does so in line with predictions of distributional learning models. Initial evidence that speaks to this question comes from research on lexically- or visually-guided perceptual learning (Norris et al., 2003; **bertelson20023?**; **kraljic-samuel2005?**). In these paradigms, listeners are exposed to phonetically manipulated instances of a sound category (e.g., making the "s" in "embassy" sound almost like an "sh"), mixed with many filler words without that sound. Following such exposure, listeners are known to shift their categorization function. For example, after being exposed to instances of "sh"-like "s" listeners categorize more tokens along the "s"-"sh" continuum as "s". Recent work within those paradigms has found that the magnitude of the boundary shift increases for listeners who are exposed to more instances of the shifted sound Kleinschmidt & Jaeger (2011). This suggest that adaptation accumulates with exposure, rather than being an all or nothing process (**cummings2023?**). There are, however, important limitations to these findings. Perceptual recalibration paradigms, at least as used traditionally, limit experimenters' control over the phonetic properties of the exposure stimuli: shifted sound instances are selected to be perceptually ambiguous (e.g., between "s" and "sh"), not to exhibit specific phonetic distributions. To the extent that researchers have aimed to understand the consequences of phonetic properties on the degree of boundary shift following exposure, this has been limited to post-hoc analyses (**drouin2016?**; **kaljic-samuel2007?**; **other-cummings?**). It is thus an open question to what extent the boundary shifts observed in such experiments reflect not only the quantity, but also the distribution of phonetic properties, during exposure (as predicted by distributional learning models).

This motivates the present study. We modify the distributional learning paradigm of Clayards et al. (2008) to shed light on the cumulative effects of incremental adaptation. We exposure participants to instances of "d" and "t", and manipulate the distribution of VOT between participants, while intermittently testing within-participants how listeners' categorization functions change with exposure. The resulting repeated exposure=test design is shown in Figure 1.

The use of repeated testing deviates from previous work, and not without challenges.

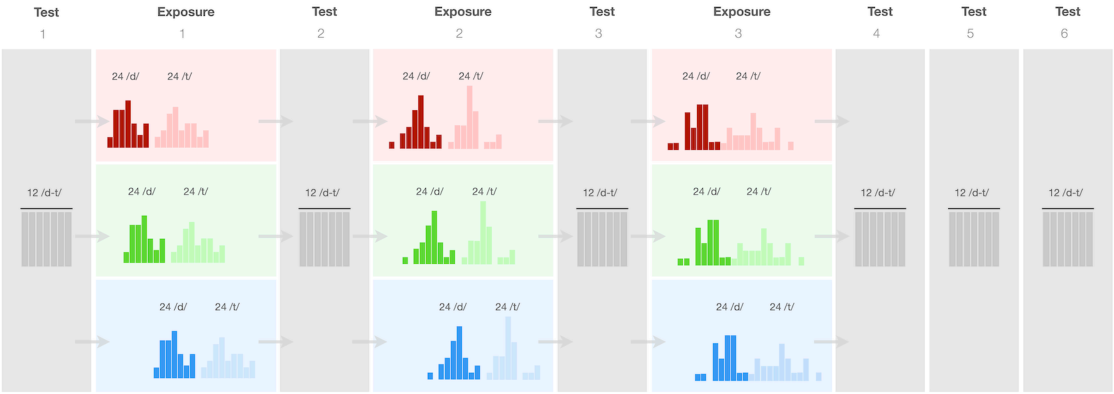


Figure 1. Exposure-test design of the experiment. Test blocks presented identical stimuli within and across conditions

Previous work has instead employed ‘batch testing’ designs, in which changes in categorization responses are assessed only after extended exposure to hundreds of trials or by averaging over similarly extended exposure (Clayards et al., 2008; Harmon et al., 2019; Idemaru & Holt, 2011, 2020; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016; Munson, 2011; Nixon et al., 2016; Theodore & Monto, 2019). By introducing intermittent testing we aim to assess how increasing exposure affects listeners’ perception without making strong assumptions about the nature of these changes (such as assuming linearity, or penalizing non-linearity, of changes over trials). However, we cannot afford *extended* intermittent testing for three reasons. First, listeners’ attention span is limited. Even prior to additional testing, typical distributional learning experiments span 200-400+ trials. Extending them further risks increasing attentional lapses and deteriorating data quality. Second, previous work has found that repeated testing over uniform test continua can reduce or undo the effects of informative exposure (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019). Third, holding the distribution of test stimuli constant across exposure condition inevitably means that the relative unexpectedness of these test stimuli differs between the exposure conditions By keeping tests short relative exposure (12 vs. 48 trials), we aimed to minimize the influence of test trials on adaptation. The final three 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.

We also made several additional adjustments to the paradigms used in previous work, meant to increase the ecologically validity of both stimuli and exposure distributions. This serves the longer-term goal of bridging the gap between research paradigms that afford control over phonetic properties at the cost of ecological validity, and paradigms that afford high ecological validity (e.g. adaptation to natural accents) at the cost of control. We describe the adjustments in more detail under Methods but briefly anticipate them here. The pioneering works we build on employed speech stimuli that were clearly identifiable as synthesized, sounding robotic, and did not exhibit natural correlations between phonetic cues (Clayards et al., 2008; Kleinschmidt & Jaeger, 2016). We instead created natural sounding stimuli (building on Theodore & Monto, 2019) that exhibited correlations between VOT and other cues to word-initial "d"-"t" that typical to everyday speech (**REF?**). Previous work also *designed* rather than *sampled* exposure distributions. As a consequence, exposure distributions in these experiments were symmetrically balanced around the category means [see also Harmon et al. (2019); Idemaru and Holt (2011); Idemaru and Holt (2020); Vroomen et al. (2007); a.o.]—unlike in everyday speech input which constitutes heterogeneous *random samples* of the underlying phonetic distributions. Indeed, all previous studied we build on exposed listeners to categories with *identical* variances (e.g., identical variance along VOT for /b/ and /p/, Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; or /g/ and /k/, Theodore & Monto, 2019). This, too, is highly atypical for everyday speech input (Lisker & Abramson, 1964). We instead expose listeners to random samples of phonetic cues that exhibit natural asymmetries in category variance based on a phonetically annotated database of word-initial /d/ and /t/ in US English (Chodroff & Wilson, 2018).

1.1 Other notes

The predominant paradigms in research on adaptive speech perception are, however, not well-suited to address this question. As Cummings and Theodore (2023) summarize, "most research […] has focused on identifying the conditions that are necessary for adaptation to occur" and "consistent with [this goal], outcomes […] are most often considered as a binary result—does any learning occur, or not?" As a consequence, much remains unknown about how exposure comes to affect perception. It is unclear, for example, whether adaptive changes accumulate depending on both the amount of speech input and its statistical properties in the way predicted by the most explicit theoretical frameworks (e.g., the ideal adaptor, Kleinschmidt & Jaeger, 2015; C-CuRE, Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011).

Typical paradigms manipulate exposure between listeners, and then assess the effects of exposure on subsequent test stimuli that are identical for all groups (Schertz & Clare, 2020; for review, see **baese-berk2018?**). These types of paradigms have provided evidence that adaptation to an unfamiliar talker can be rapid. For example, a thought-provoking finding by Clarke and Garrett (2004) suggests that exposure to eighteen sentences from an L2-accented talker—less than two minutes of speech—can be sufficient to facilitate significantly faster processing of that speech. This finding has since been replicated and extended to show that equally short exposure can facilitate recognition that is both faster and more accurate (Xie, Weatherholtz, et al., 2018; Xie, Liu, & Jaeger, 2021; for related results, see also **bradlow2023?**; **xie2017?**). Other work has traded the ecological validity of natural L2 accents against increased control over the phonetic properties of exposure and test stimuli—a critical step towards stronger tests, as competing hypotheses about the mechanisms underlying adaptive speech perception require strong linking hypotheses mapping the acoustic input onto listeners' responses (Xie et al., 2023; **martin2023?**). One such paradigm is lexically- or visually-guided perceptual recalibration (Norris et al., 2003; **bertelson20023?**; **kraljic-samuel2005?**), in which listeners are exposed to phonetically manipulated instances of a sound (e.g., making the "s" in "embassy" sound almost like an "sh"), mixed with many filler words without that sound. Following such exposure, listeners are known to shift their categorization function, so as to categorize more tokens along the "s"-"sh" continuum as "s". Recent work within those paradigms has found that as little as four phonetically shifted instances of a sound category can be sufficient to significantly alter listeners' categorization boundary (Liu & Jaeger, 2018, 2019; Vroomen et al., 2007; **cummings2023?**). The same studies have found that exposure seems to accumulate, leading to larger boundary shifts for listeners who were exposed to more instances of the shifted sound (up to a point, Liu & Jaeger, 2018; Vroomen et al., 2007; see also Kleinschmidt & Jaeger, 2011; **kleinschmidt-jaeger2012?**). Findings like these suggest that even rapid adaptation can be cumulative, rather than being an all or nothing process.

There are, however, important limitations to what perceptual recalibration paradigms can tell us about incremental adaptation. As is typical for such paradigms, all of the above experiments exposed listeners to shifted pronunciations that were always lexically or visually labeled stimuli (e.g., embedding the "sh"-like "s" in the word "embassy", which effectively labels it as an "s"). Such labeling is known to facilitate adaptation (**burchill2018?**; **burchill2023?**)—indeed, if shifted pronunciations are embedded in minimal pair or nonce-word context, listeners do no longer shift their categorization boundary (Norris et al., 2003; **REF-theodore?**). In everyday speech perception, however, listeners often have uncertainty about the word they are hearing, and must either use contextual information to label the input or adapt from unlabeled input. Perceptual recalibration paradigms, at least as used traditionally, also limit experimenters' control over the phonetic properties of the exposure stimuli: shifted sound instances are selected to sound ambiguous between, e.g., "s" and "sh", not based on their phonetic properties. To the extent that researchers have aimed to understand the consequences of those phonetic properties on the degree of boundary shift following exposure, this has involved post-hoc analyses (**drouin2016?**; **other-cummings?**). It is thus an open question to what extent the boundary shifts observed in such experiments reflect not only the quantity, but also the distribution of phonetic properties, during exposure [as would be expected under, e.g., the ideal adaptor framework].

The present work thus employs a novel repeated-exposure-test paradigm that explicitly control the distribution of phonetic properties during exposure. [clayards, bejjanki; kj16, k20; see also theodore-monto2019]

[existing evidence comes from paradigms that emphasize ecological validity at the cost of less control: accent adaptation. Recent work expands on these findings using PR. Here we]

for distributional only exposure-test or exposure/test

Contributions: + ecological validity of stimuli: + how they sound + correlation of VOT and f0 + anti-correlation of VOT and vowel duration + ecological validity of distribution + variances that mimic those found in natural speech + samples drawn from theoretical distributions, rather than presentation of perfectly symmetrical stimulus samples

Possible framings: 1a) mechanisms remain unknown 1b) need for stronger tests of theories 2) incremental cumulative effects

1.2 Maryann's most recent intro

Recent reviews have identified distributional learning of marginal cue statistics ('normalization,' Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011; **magnuson-nusbaum2007?**) or the statistics of cue-to-category mappings as an important mechanism affording this adaptivity ('representational learning,' Clayards et al., 2008; Davis & Sohoglu, 2020; Idemaru & Holt, 2011; Kleinschmidt & Jaeger, 2015; for review, Schertz & Clare, 2020; Xie et al., 2023). This hypothesis has gained considerable influence over the past decade, with findings that changes in listener perception are qualitatively predicted by the statistics of exposure stimuli (Bejjanki et al., 2011; Clayards et al., 2008; Idemaru & Holt, 2020; Kleinschmidt & Jaeger, 2012; Munson, 2011; Nixon et al., 2016; Tan et al., 2021; Theodore & Monto, 2019; for important caveats, see Harmon et al., 2019).

Viewing speech perception as an adaptive process has been pivotal in our understanding of how human listeners overcome the lack of invariance problem; a problem fully appreciated when one begins to map out the variability of acoustic-phonetic cues that point to a single linguistic category (e.g. Delattre, Liberman, & Cooper, 1955; Newman, Clouse, & Burnham, 2001; Peterson & Barney, 1952); compounded when talker sex, age, social class, dialect and a host of other contexts are factored into consideration. Listeners' aptitude at speech comprehension however, belie this challenge. Given the uncertainty involved it is not surprising models of spoken word recognition that allow for probabilistic outcomes have left a lasting impression (Norris & McQueen, 2008; **mclleland-elman1986?**; **vitevitch-luce?**).

Over the past 20 years there have been prolific investigations into how and when listeners adjust their phonological categories after hearing acoustically manipulated speech sounds. These manipulations take place at the margins of linguistic categories where perception can be heavily influenced by the contexts in which they are presented (McQueen, Cutler, & Norris, 2006; Norris et al., 2003). A sound that is ambiguous between /s/ and /sh/ presented in the utterance *contradiction* would bias its interpretation as /sh/ since *contradicson* is not a word. Repeated exposure to the sound in such biasing word contexts reliably elicits a shift in perception along the /s/-/sh/ continuum in subsequent testing – those having heard the sound in /sh/-biasing words tend to give more /sh/ responses; vice-versa for those who were exposed to it in /s/-contexts.

This perceptual recalibration of less prototypical category members has also been induced under audio-visual manipulations (Bertelson et al., 2003; Vroomen et al., 2007). The paradigm has been exploited to its fullest to investigate, among other things, the sustainability of perceptual changes (**eisner-mcqueen2006?**; **kraljic-samuel2005?**), its generalizability to members of the same phonological class (**kraljic-samuel2006?**), and its generalizability to other talkers (Reinisch & Holt, 2014; **kraljic-samuel2007?**).

In general, these findings are compatible with exemplar and other probabilistic updating frameworks that link the distributions of cues to changes in category mappings hence perceptual recalibration findings can to an extent inform general understanding of talker adaptation. But the mechanisms that underlie the perceptual changes observed are still not well understood and therefore remain a point of debate. Some positions remain less specified than others. For instance the proposal that listeners expand their categories when confronted with unfamiliar accents or that they "relax their criteria" for category membership (Zheng and Samuel (2020); (**schmale2012?**); (**floccia2006?**); (**bent2016?**)). While it is possible that apparent perceptual shifts post-exposure can be explained by processes independent of distributional learning (**clarke-davidson2008?**; see Xie et al., 2023 for simulations) what is needed are better specified hypotheses coupled with stronger predictions and tests to weigh the evidence (Schertz & Clare, 2020; Xie et al., 2023; **bent-baese-berk2021?**).

Analytic frameworks that facilitate modelling of perceptual processes conditioned on different assumptions offer a way forward. If robust speech recognition involves learning from the input under varying contexts in a rational manner, it has to account for the implicit assumptions that listeners seem to bring to any speech perception task with regard to cue-category mappings, and be able to explain how they reconcile these assumptions with recent input. Theories that explicitly bring this to bear include the influential exemplar models (Apfelbaum & McMurray, 2015; Pierrehumbert, 2001; **johnson1996?**), Bayesian inference models (Hitczenko & Feldman, 2016; Kleinschmidt & Jaeger, 2015; Kronrod, Coppess, & Feldman, 2016; **feldman2009?**), and error-driven learning (Harmon et al., 2019).

In a recent example Cummings and Theodore (2023) working within the ideal adaptor framework, predicted that perceptual recalibration could have graded effects. This logic follows from the general premise that adaptation is the outcome of weighted updates of listener prior expectations of cue-category mappings with the statistics of talker input. By manipulating the number of times an ambiguous sound between /s/ and /sh/ was heard between participants and within each biasing context (1, 4, 10 or 20 occurences) they showed that the size of the putative perceptual recalibration effect correlated with the frequency of the ambiguous tokens. Model simulations qualitatively predicted behavioral results and provided strong evidence of a mechanism that is sensitive to cue statistics. This result corroborates earlier modelling efforts of Kleinschmidt and Jaeger (2011) which demonstrated that incremental bayesian belief-updating is a possible mechanism behind what has been believed to be dichotomous perceptual phenomena – selective adaptation and perceptual recalibration.

The present study was devised in similar spirit to past studies guided by an understanding of language as inference and learning under uncertain conditions (Clayards et al., 2008; Kleinschmidt & Jaeger, 2011, 2016; **fine2010?**). In particular we aim to subject the hypothesis that talker adaptation results from distributional learning with incremental belief updating to a stronger test. While studies of perceptual recalibration that demonstrate graded learning effects based on the quantity of evidence support this hypothesis, there are limitations to the paradigm that preclude deeper investigation. Talker-specific learning involves inferring the means and variances of her cue-category mappings. This task is made more difficult for talkers with extreme cue shifts that fall beyond the prior expectations of listeners because an entire remapping of the cue space is required (Sumner, 2011). In perceptual recalibration listeners are presented with maximally informative instances of the same ambiguous acoustic-phonetic token essentially providing ideal but very unnatural circumstances for learning to occur. However even this has a limit – exposure to a certain number of critical trials (about 20 trials in lexical context studies (**cummings-theodore2022?**; **tzeng2021?**); 64 trials in audio-visual context studies(Vroomen et al., 2007)) – do not bring additive learning effects.

Here we build on the pioneering work of Clayards et al. (2008); Kleinschmidt and Jaeger (2016); Theodore and Monto (2019); Kleinschmidt (2020) with some design innovations that we believe affords a productive test of the core claims of an ideal adaptor account of speech perception. In Kleinschmidt and Jaeger (2016) L1-US English listeners heard recordings of

/b/-/p/ minimal pair words like *beach* and *peach* that were acoustically manipulated. Separate groups of listeners were exposed to different distributions of voice onset times (VOTs)—the primary cue distinguishing word-initial voicing —that were shifted by up to +30 ms, relative to what one might expect from a 'typical' talker (Figure 2A). In line with the distributional learning hypothesis, listeners' category boundary or point of subjective equality (PSE)—i.e., the VOT for which listeners are equally likely to respond "b" or "p"—shifted in the same direction as the exposure distribution (Figure 2B). Kleinschmidt and Jaeger (2016) and closely related work have been able to show perceptual shifts move qualitatively in the direction of the manipulated distributions but so far none of them were designed to test incremental adaptation. We propose to fill that gap with a novel test-exposure-test design. In doing so we aim to estimate listeners prior expectations about the category mappings for our test talker before they receive further informative exposure and to document how quickly, from the onset of exposure, does the distributional learning effect emerge. The latter point is something that remains opaque in previous work because of the lack of test blocks. Given the substantial evidence that adaptation is rapid (e.g. under 5 mins in L2 accent adaptation; 4-10 trials in perceptual recalibration) listeners may show learning effects very early on in distributional learning as well. On the other hand, given the comparatively more naturalistic task of inferring talker distributions over a range of cues, learning effects may take longer to show.

In experimental work researchers often have to consider the generalizability of their results which leads to questions about ecological validity. There is a trade-off between ecological validity of the experimental design and the desired degree of control over the variables. Questions about ecological validity of prior work in distributional learning pertain to two features. First, the stimuli which were generated with a synthesiser, had an obvious machine-like quality(Clayards et al., 2008; Kleinschmidt & Jaeger, 2016). Second, the pairs of distributions of voiced and voiceless categories were always identical in their variances (see also Theodore & Monto, 2019) which adds to the artificiality of the experiment. In our description of methods below we show how we can begin to improve on these features through the stimuli and the setting of exposure conditions.

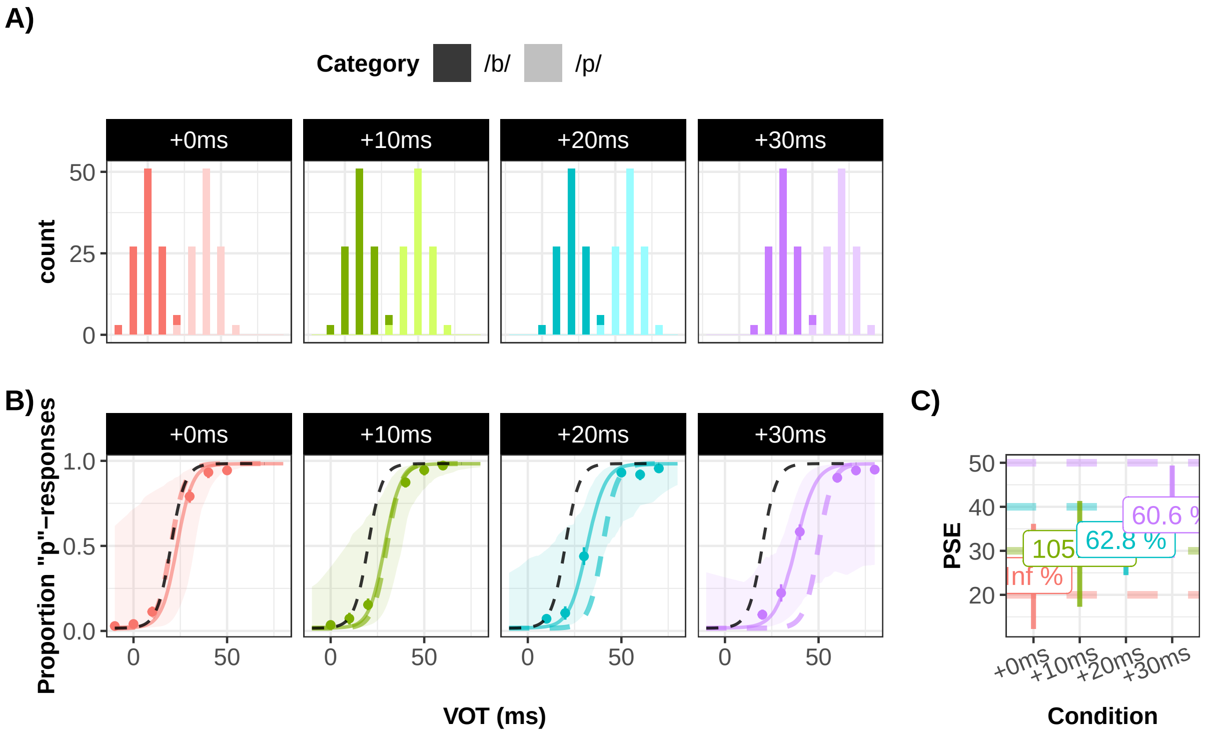


Figure 2. Design and results of Kleinschmidt and Jaeger (2016) replotted. **Panel A:** Different groups of participants were exposed to different shifts in the mean VOT of /b/ and /p/. **Panel B:** categorization functions fitted to the last 1/6th of all trials depending on the exposure condition (shift in VOT means of /b/ and /p/). For reference, the black dashed line shows the categorization function of the 0-shift condition. The colored dashed lines shows the categorization function expected for an ideal observer that has fully learned the exposure distributions. **Panel C:** Mean and 95% CI of participants' points of subjective equality (PSEs), relative to the PSE of the ideal observers.

1.3 Previous intro

For example, influential models of adaptive speech perception predict proportional, rather than sublinear, shifts (for proof, see SI **??**). This is the case both for incremental Bayesian belief-updating model (Kleinschmidt & Jaeger, 2011) and general purpose normalization accounts (McMurray & Jongman, 2011)—models that have been found to explain listeners' behavior well in experiments with less substantial changes in exposure. There are, however, proposals that can accommodate this finding. Some proposals distinguish between two types of mechanisms that might underlie representational changes, *model learning* and *model selection* (Xie, Weatherholtz, et al., 2018, p. 229). The former refers to the learning of a new category representations—for example, learning a new generative model for the talker (Kleinschmidt & Jaeger, 2015, pt. II) or storage of new talker-specific exemplars (Johnson, 1997; Sumner, 2011). Xie and colleagues hypothesized that this process might be much slower than is often assumed in the literature, potentially requiring multiple days of exposure and memory consolidation during sleep (see also Fenn & Hambrick, 2013; Tamminen, Davis, Merkx, & Rastle, 2012; Xie, Earle, & Myers, 2018). Rapid adaptation that occurs within minutes of exposure might instead be achieved by selecting between *existing* talker-specific representations that were learned from previous speech input—e.g., previously learned talker-specific generative models (see mixture model in Kleinschmidt & Jaeger, 2015, pp. 180–181) or previously stored exemplars from other talkers (Johnson, 1997). Model learning and model selection both offer explanations for the sublinear effects observed in Kleinschmidt and Jaeger (2016). But they suggest different predictions for the evolution of this effect over the course of exposure.

Under the hypothesis of model learning, sublinear shifts in PSEs can be explained by assuming a hierarchical prior over talker-specific generative models ($p(\Theta)$ in Kleinschmidt & Jaeger, 2015, p. 180). This prior would 'shrink' adaptation towards listeners' priors—similar to the effect of random by-subject or by-item effects in generalized linear mixed-effect models, which shrink group-level effect estimates towards the population mean of the data (Baayen, Davidson, & Bates, 2008). Critically, as long as these priors attribute non-zero probability to even extreme shifts (e.g., the type of Gaussian prior used in mixed-effects models), this predicts listeners' PSEs will continue to change with increasing exposure until they have converged against the PSE that is ideal for the exposure statistics. In contrast, the hypothesis of model selection predicts that rapid adaptation is more strictly constrained by previous experience: listeners can only adapt their categorization functions up to a point that corresponds to (a mixture of) previously learned talker-specific generative models. This would imply that at least the earliest moments of adaptation are subject to a hard limit (Figure 3): exposure helps listeners to adapt their interpretation to more closely aligned with the statistics of the input, but only to a certain point.

The present study employs a novel incremental exposure-test paradigm to address two questions. We test whether the sublinear effects of exposure observed in recent work replicate for exposure that (somewhat) more closely resembles the type of speech input listeners receive on a

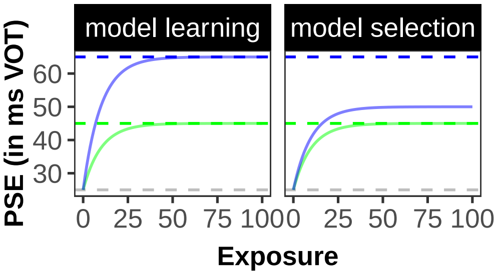


Figure 3. Contrasting predictions of model learning and model selection hypotheses about the incremental effects of exposure on listeners' categorization function. Both hypothesis predict incremental adaptation towards the statistics of the input, as well as constraints on this adaptation. The two hypotheses differ, however, in that model selection predicts a hard limit on how far listeners' can adapt during initial encounters with an unfamiliar talker.

daily basis. And, we evaluate the predictions of the model learning and selection hypotheses against human perception. We take this question to be of interest beyond the specific hypotheses we contrast: whether there are hard limits to the benefits of exposure to unfamiliar speech patterns ultimately has consequences for education and medical treatment.

All data and code for this article can be downloaded from https://osf.io/hxcy4/. The 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, 2022; RStudio Team, 2020), while changing any of the parameters of our models (see SI, **??**).

2 Experiment

2.1 Methods

2.1.1 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 be English speaking monolinguals, and (3) had not previously participated in any experiment from our lab on Prolific. Prior to the start of the experiment, participants had to confirm that they (4) had spent the first 10 years of their life in the US, (5) were in a quiet place and free from distractions, and (6) wore in-ear or over-the-ears headphones that cost at least $15. An additional

115 participants loaded the experiment but did not start or complete it.[1]

Participants took an average of 31.6 minutes to complete the experiment (SD = 20 minutes) 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, S, 2020).

2.1.2 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. In addition to these critical minimal pairs we also recorded three words that did not did not contain any stop consonant sounds ("flare", "share", and "rare"). These word recordings were used for catch trials. Stimulus intensity was normalized to 70 dB sound pressure level for all recordings.

The critical minimal pair recordings were used to create four VOT continua using a script (Winn, 2020) in Praat (Boersma & Weenink, 2022). This approach resulted in continuum steps that sound natural (unlike the highly robotic-sounding stimuli employed in Clayards et al., 2008; Kleinschmidt & Jaeger, 2016). A post-experiment survey asked participants: "*Did you notice anything in particular about how the speaker pronounced the different words (e.g. till, dill, etc.)?*" No participant reported that the stimuli sounded unnatural. 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

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

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, **??**).

The VOTs generated for each continuum ranged from -100 to +130 ms in 5 ms steps.[2] A norming experiment (N = 24 participants) reported in the SI (**??**) was used to select the three minimal pair continua that elicited the most similar categorization responses (*dill-till*, *din-tin*, and *dip-tip*). These three continua were used to create the exposure conditions shown in Figure 1.

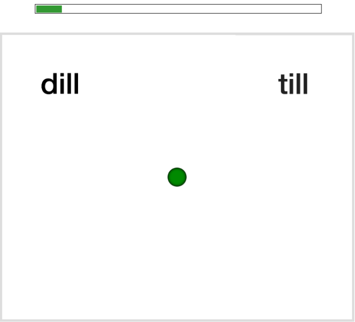
2.1.3 Procedure

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 (Woods, Siegel, Traer, & McDermott, 2017), and were instructed to not change the volume throughout the experiment. Following instructions, participants completed 234 two-alternative forced-choice categorization trials (Figure 4). 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.

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

[2] We follow previous work (Kleinschmidt, 2020; Lisker & Abramson, 1964) and refer to pre-voicing as negative VOTs though we note that pre-voicing is perhaps better conceived of as a separate phonetic feature (for discussion, see **REF?**). Estimates of the proportion of voiced stops produced with pre-voicing in L1-US English vary substantially between studies (between 20% and 57%) (Dmitrieva, Llanos, Shultz, & Francis, 2015; e.g. Lisker & Abramson, 1967; Smith, 1978; Westbury, 1979). Because pre-voicing is not regarded as a phonemic determinant of English, some studies either discard such data or ignore them altogether (e.g. Zue (1976); Klatt (1975); Chodroff and Wilson (2017)). In some studies that do report pre-voicing, the majority of the tokens were attributed to a minority of talkers (Flege & Brown Jr, 1982; e.g. Lisker & Abramson, 1967). Although speakers tend to prefer one type of production over the other they do not typically use one type exclusively (Docherty, 2011).



dill

till

Figure 4. Example trial display. When the green button turned bright green, participants had to click on it to play the recording.

and across conditions, always including 12 minimal pair trials assessing participants' categorization at 12 different VOTs (-5, 5, 15, 25, 30, 35, 40, 45, 50, 55, 65, 70 ms). A uniform distribution over VOTs was chosen to maximize the statistical power to determine participants' categorization function. The assignment of VOTs to minimal pair continua was randomized for each participant, while counter-balancing it within and across test blocks. Each minimal pair appear equally often within each test block (four times), and each minimal pair appear with each VOT equally often (twice) across all six test blocks (and no more than once per test block).

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 4. 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.

Exposure blocks. Each exposure block consisted of 24 /d/ and 24 /t/ trials, as well as 6 catch trials that served as a check on participant attention throughout the experiment (2 instances for each of three combinations of the three catch recordings). With a total of 144 trials, exposure was substantially shorter than in similar previous experiments (cf. 228 trials in Clayards et al., 2008; 222 trials in Kleinschmidt, 2020; 2 x 236 trials, Theodore & Monto, 2019; 456 trials, Nixon et al., 2016).

The distribution of VOTs across the 48 /d/-/t/ trials depended on the exposure condition.

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 at 5 ms and 8.9 ms, respectively. The mean and standard deviations for /t/ were set at 50 ms and 16 ms, respectively. To create more realistic VOT distributions, we *sampled* from the intended VOT distribution (top row of Figure 5). 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., 2008; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016).

Half of the /d/ and half of the /t/ trials were labeled, the other half was unlabeled. Earlier distributional learning studies have mostly used fully unlabeled exposure (Bejjanki et al., 2011; Clayards et al., 2008; **nixon?**). This contrasts with visually- or lexically-guided perceptual learning studies, which use labeled exposure (Bertelson et al., 2003; Norris et al., 2003; Vroomen et al., 2007; **kraljic-samuel2005?**). Such labeling is known to facilitate adaptation (**burchill2018?**; **burchill2023?**; but see Kleinschmidt, Raizada, & Jaeger, 2015)—indeed, if shifted pronunciations are embedded in minimal pair or nonce-word context, listeners do no longer shift their categorization boundary (Norris et al., 2003; **REF-theodore?**; **babel?**). While lexical contexts often disambiguate sounds in everyday speech, that is not *always* the case: especially, when confronted with unfamiliar accents, listeners often have uncertainty about the word they are hearing, and must either use contextual information to label the input or adapt from unlabeled input. Here, we thus aimed to strike a compromise between always and never labeling the input (paralleling one of the conditions in Kleinschmidt et al., 2015).

Unlabeled trials were identical to test trials except that the distribution of VOTs across those trials was bimodal (rather than uniform), and determined by the exposure condition.[3]

[3] Previous studies have estimated changes in participants' categorization responses by analyzing responses on unlabeled exposure trials (e.g., Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; Theodore & Monto, 2019). This approach compares responses across different values of acoustic-phonetic cues (since the exposure inputs differed by exposure condition), so that assumptions baked into the analysis approach (e.g., linearity along the acoustic-phonetic continuum) can potentially bias the results. Here we avoid this issue by holding test stimuli constant (see also Kleinschmidt, 2020, Experiment 4).

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 ms (see Figure 5). This approach exposes participants to heterogeneous approximations of normally distributed VOTs for /d/ and /t/ that varied across blocks, while holding all aspects of the input 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 3 (exposure condition) x 3 (block order) x 2 (placement of response options) lists.

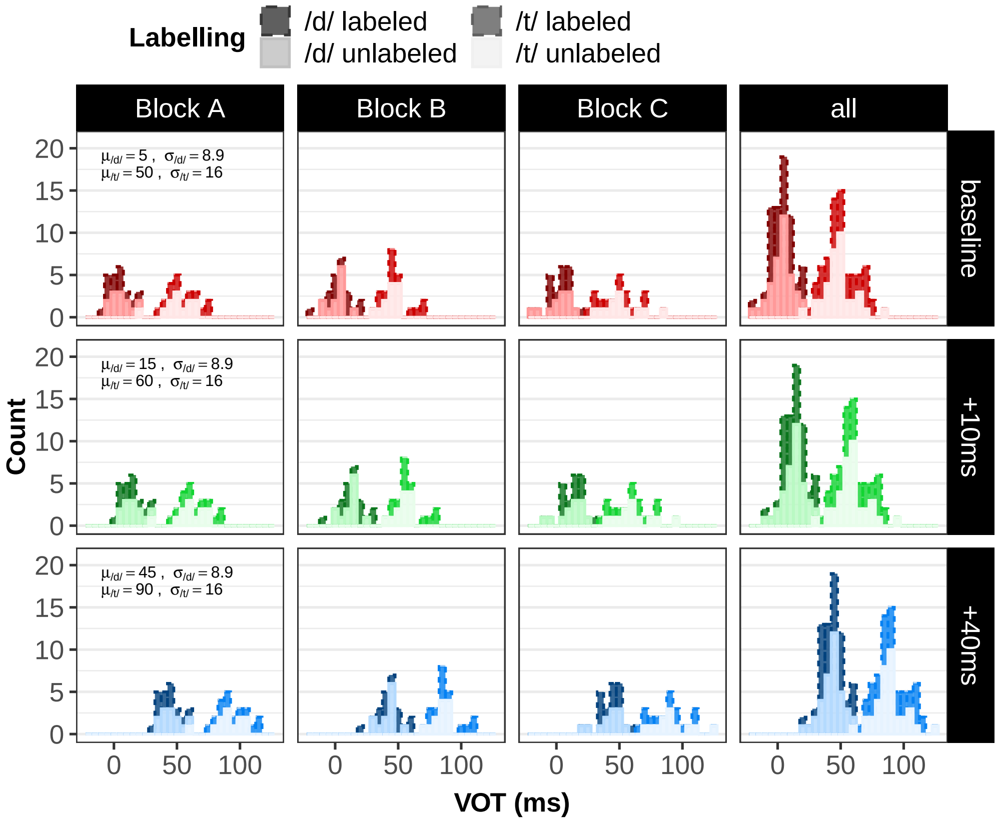


Figure 5. Histogram of voice onset times (VOTs) for each of the three exposure blocks A-C by trial type (/d/ or /t/, labeled or unlabeled) and exposure condition (baseline vs. +10 vs. +40). Each exposure block contained 12 labeled /d/, 12 labeled /t/, 12 unlabeled /d/, and 12 unlabeled /t/ trials, as well as 6 catch trials (not shown). Except for the shift in VOTs (+0, 10 or 40 ms VOT to each trial), the VOT distribution of these trials was identical across exposure conditions. The order of exposure blocks A-C was counter-balanced across participants using a Latin-square design.

2.1.4 Exclusions

Due to data transfer errors 4 participants' data were not stored and therefore excluded from analysis. We further excluded from analysis participants who committed more than 3 errors out of the 18 catch trials (<83% accuracy, N = 1), participants who committed more than 4 errors out of the 72 labelled trials (<94% accuracy, N = 0), participants with an average reaction time more than three standard deviations from the mean of the by-participant means (N = 0), participants who had atypical categorization functions at the start of the experiment (N = 2, see SI, **??** for details), and participants who reported not to have used headphones (N = 0). This left for analysis 17,136 exposure and 8,568 test observations from 1,071 participants (94% of total), evenly split across the three exposure conditions.

“‘{r-remove-unused-objects}

```
## Results
```

```
We analyzed participants' categorization responses during exposure and test blocks in two sepa
```

Each psychometric model regressed participants' categorization responses against the full facto

We begin by presenting the overall effects, averaging across all test blocks. This part of our

2.2 [1] “VOT test mean: 35.8333333333333”

2.3 [1] “VOT test mean: 35.8333333333333”

2.4 [1] “VOT test mean:”

2.5 [1] “VOT test mean:”

“‘

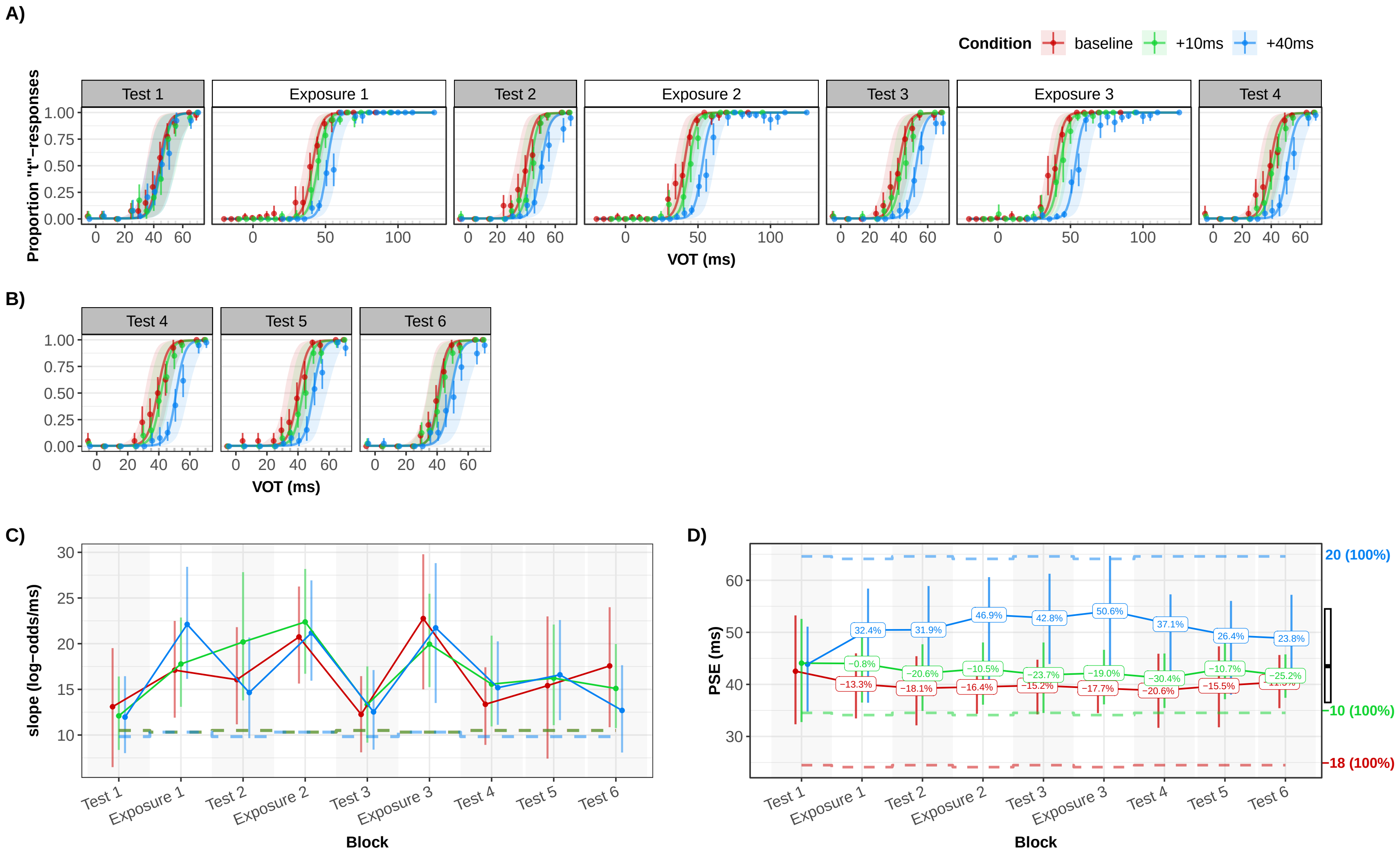


Figure 6. Summary of results. **Panel A:** Changes in listeners psychometric categorization functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks (only unlabeled trials were included in the analysis of exposure blocks since labeled trials provide no information about listeners' categorization function). Point ranges indicate the mean proportion of "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 categorization functions shown in Panels A-B. Point ranges represent the posterior medians and their 95% CI. Dashed reference lines show the intercepts and PSEs that naive (non-rational) learner would be expected to converge against after sufficient exposure (an ideal observer model that knows the exposure distributions). Percentage labels indicate the amount of shift

2.5.1 Does exposure affect participants' categorizations (averaging across all blocks)?

We first used the psychometric mixed-effects model to assess whether the exposure conditions had the expected effects across all test blocks *relative to each other*. Unsurprisingly, participants were more likely to respond "t" the larger the VOT ($\hat{\beta} = 15.09,\ 90\%\text{−CI} = [12.377, 17.625],\ BF = Inf,\ p_{posterior} = 1$). Critically, exposure affects participants' categorization responses in the expected direction. Marginalizing across all blocks, participants in the +40 condition were less likely to respond "t" than participants in the +10 condition ($\hat{\beta} = -2.26,\ 90\%\text{−CI} = [-3.258, -1.228],\ BF = 162.3,\ p_{posterior} = 0.994$) or the baseline condition ($\hat{\beta} = -3.08,\ 90\%\text{−CI} = [-4.403, -1.669],\ BF = 215.2,\ p_{posterior} = 0.995$). There was also evidence—albeit less decisive—that participants in the +10 condition were less likely to respond "t" than participants in the baseline condition ($\hat{\beta} = -0.82,\ 90\%\text{−CI} = [-1.887, 0.282],\ BF = 8.9,\ p_{posterior} = 0.899$). That is, the +10 and +40 conditions resulted in categorization functions that were shifted rightwards compared to the baseline condition, as also visible in Figures 6.

This replicates previous findings that exposure to changed VOT distributions changes listeners' categorization responses (for /b/-/p/: Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016; for /g/-/k/, Theodore & Monto, 2019). Having established that exposure affected categorization, we turn to the questions of primary interest. Incremental changes in participants' categorization responses can be assessed from three mutually complementing perspectives. First, we compare how exposure affects listeners' categorization responses relative to other exposure conditions. This tests how early in the experiment differences between exposure conditions began to emerge. Second, we compare how exposure affects listeners' categorization responses within each condition relative to listeners' responses prior to any exposure. This assesses how the exposure conditions relate to participants' prior expectations. Most importantly, however, it tests the subtly different predictions of the model learning and selection hypotheses—whether changes in listeners' categorization responses are strongly constrained. Third and finally, we compare changes in listeners' responses to those expected from an ideal observer that has fully learned the exposure distributions. This tests whether the sublinear effects observed in Kleinschmidt and Jaeger (2016) replicate in our repeated exposure-test paradigm with the improvements the present study makes to ecological validity.

2.5.2 Comparing across exposure conditions: How quickly does exposure begin to affect participants' responses?

Figure 6A suggests that differences between exposure conditions emerged early in the experiment: already in Test 2, listener's categorization functions seem to be shifted rightwards (larger PSEs) in the +40 condition compared to the +10 condition, and in the +10 condition compared to the baseline condition. This is confirmed by the Bayesian hypothesis tests summarized in Table 1. Prior to any exposure, during Test 1, participants' responses did not differ across exposure condition (all BFs > XXX). After exposure to only 24 /d/ and 24 /t/ stimuli, during Test 2, participants' responses differed between exposure conditions (BFs > 13.7). The difference between the +40 condition and the +10 or baseline condition kept increasing with exposure up to Test 4. Additional hypothesis tests in Table 2 show that the change from Test 1 to 2 was largest (BF = 57.82), followed by the change from Test 2 to 3 (BF = 10), with only minimal changes from Test 3 to 4 (BF = 1.68). Qualitatively paralleling the changes across blocks for the +40 condition, the change in the difference between the +10 and baseline conditions was largest from Test 1 to 2 (BF = 5.42), and then somewhat decreased from Test 2 to Test 4 (BFs < 1). The comparison across exposure conditions thus suggests that changes in listeners' categorization responses emerged quickly—indeed, they were present already *during* the first exposure block (see SI, **??**)—but then leveled off. The comparison across exposure conditions also yields one result that is, at first blush, surprising: while the difference between the +10 and the baseline condition emerged already after the first exposure block, this difference *de*creased, rather than increased, with additional exposure from Test 2 to 3 (see second row of Table 2). We return to this effect below.

Tables 1 and 2 also reveal the consequences of repeated testing. The difference between exposure conditions decreased from Test 4 to 6 (BFs > 4.3; see also Figure 6B & D). On the final test block, the +10 condition did not differ any longer from the baseline condition. Only the differences between the +40 condition relative to the +10 and baseline conditions persisted, albeit substantially reduced compared to Test 4. This pattern of results replicates previous findings that repeated testing over uniform test continua can undo the effects of exposure (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019), and extends them from perceptual recalibration paradigms to distributional learning paradigms (see also Kleinschmidt, 2020). One important methodological consequence of these findings is that longer test phases do not necessarily increase the statistical power to detect effects of adaptation (unless analyses take the effects of repeated testing into account, as in the approach developed in Liu & Jaeger, 2018). Analyses that average across all test tokens—as remains the norm—are bound to systematically underestimate the adaptivity of human speech perception.

Table 1

When did exposure begin to affect participants' categorization responses? When, if ever, were these changes undone with repeated testing? This table summarizes the simple effects of the exposure conditions for each test block.

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Test block 1 (pre-exposure)					
+10 vs. baseline = 0	-0.34	0.75	[-2.025, 1.437]	3.3	0.77
+40 vs. +10 = 0	0.25	0.73	[-1.338, 1.903]	3.7	0.79
+40 vs. baseline = 0	-0.08	0.91	[-2.124, 2.082]	4.8	0.83
Test block 2					
+10 vs. baseline	-1.45	0.88	[-2.933, 0.181]	13.7	0.93
+40 vs. +10	-2.08	0.99	[-3.824, -0.173]	24.3	0.96
+40 vs. baseline	-3.49	1.24	[-5.635, -1.072]	54.2	0.98
Test block 3					
+10 vs. baseline	-0.78	0.62	[-1.888, 0.364]	7.9	0.89
+40 vs. +10	-2.80	0.82	[-4.188, -1.113]	86.0	0.99
+40 vs. baseline	-3.56	0.97	[-5.202, -1.582]	110.1	0.99
Test block 4					
+10 vs. baseline	-0.88	0.85	[-2.36, 0.847]	4.8	0.83
+40 vs. +10	-3.32	0.89	[-4.883, -1.636]	128.0	0.99
+40 vs. baseline	-4.16	1.21	[-6.275, -1.882]	122.1	0.99
Test block 5 (no additional exposure)					
+10 vs. baseline	-1.33	0.71	[-2.556, -0.003]	19.1	0.95
+40 vs. +10	-2.38	0.86	[-3.893, -0.796]	65.1	0.98
+40 vs. baseline	-3.25	1.24	[-5.307, -0.923]	53.0	0.98
Test block 6 (no additional exposure)					
+10 vs. baseline	-0.22	0.72	[-1.485, 1.114]	1.6	0.62
+40 vs. +10	-1.70	0.79	[-3.078, -0.171]	25.0	0.96
+40 vs. baseline	-2.57	1.22	[-4.58, -0.191]	24.0	0.96

Table 2

Was there incremental change from test block 1 to 4? Did these changes dissipate with repeated testing from block 4 to 6? This table summarizes the interactions between exposure condition and block, whether the differences between exposure conditions changed from test block to test block.

Hypothesis	Estimate	SE	90%-CI	BF	$p_{posterior}$
Difference in +10 vs. baseline					
Block 1 to 2: increased Δ_{PSE}	-0.85	0.78	[-2.166, 0.632]	5.42	0.84
Block 2 to 3: increased Δ_{PSE}	0.34	0.77	[-1.144, 1.761]	0.48	0.32
Block 3 to 4: increased Δ_{PSE}	0.06	0.77	[-1.382, 1.532]	0.89	0.47
Block 1 to 4: increased Δ_{PSE}	-0.42	1.26	[-2.759, 1.963]	1.70	0.63
Block 4 to 5: decreased Δ_{PSE}	-0.33	0.60	[-1.43, 0.785]	0.41	0.29
Block 5 to 6: decreased Δ_{PSE}	1.03	0.65	[-0.234, 2.164]	11.95	0.92
Block 4 to 6: decreased Δ_{PSE}	0.70	0.82	[-0.896, 2.177]	3.83	0.79
Difference in +40 vs. +10					
Block 1 to 2: increased Δ_{PSE}	-2.36	0.89	[-3.811, -0.754]	57.82	0.98
Block 2 to 3: increased Δ_{PSE}	-1.16	0.83	[-2.592, 0.312]	10.00	0.91
Block 3 to 4: increased Δ_{PSE}	-0.27	0.82	[-1.694, 1.162]	1.68	0.63
Block 1 to 4: increased Δ_{PSE}	-3.78	1.22	[-5.865, -1.447]	84.11	0.99
Block 4 to 5: decreased Δ_{PSE}	1.14	0.77	[-0.244, 2.514]	11.38	0.92
Block 5 to 6: decreased Δ_{PSE}	0.45	0.77	[-0.985, 1.787]	2.58	0.72
Block 4 to 6: decreased Δ_{PSE}	1.59	1.00	[-0.3, 3.323]	12.68	0.93
Difference in +40 vs. baseline					
Block 1 to 2: increased Δ_{PSE}	-3.16	1.02	[-4.958, -1.185]	79.00	0.99
Block 2 to 3: increased Δ_{PSE}	-0.82	1.08	[-2.749, 1.145]	3.39	0.77
Block 3 to 4: increased Δ_{PSE}	-0.20	1.08	[-2.146, 1.741]	1.34	0.57
Block 1 to 4: increased Δ_{PSE}	-4.19	1.71	[-7.219, -0.93]	45.78	0.98
Block 4 to 5: decreased Δ_{PSE}	0.80	0.92	[-0.971, 2.493]	4.16	0.81
Block 5 to 6: decreased Δ_{PSE}	1.48	0.94	[-0.36, 3.117]	10.85	0.92
Block 4 to 6: decreased Δ_{PSE}	2.27	1.27	[-0.12, 4.442]	16.47	0.94

2.5.3 Comparing within exposure conditions: How quickly does exposure begin to affect participants' responses?

Next, we compared how exposure affects listeners' categorization responses within each condition relative to listeners' responses prior to any exposure. These changes are summarized for the slope and PSE in Figure 6C & D, respectively. This visualization makes apparent two aspects of participants' behavior that were not readily apparent in the statistical comparisons we have summarized so far. First, while the PSEs for the +40 and +10 conditions were shifted rightwards compared to the baseline condition, both the +10 and the baseline condition actually shift leftwards relative to their pre-exposure starting point in Test 1. This is confirmed by Bayesian hypothesis tests summarized in Table **??**.

2.5.4 Results summary

This study was set up with several objectives in mind. We aimed to replicate previous findings on distributional learning (Kleinschmidt & Jaeger, 2016) while introducing changes to the design to a) increase the ecological validity of results b) illuminate how soon distributional learning effects can be detected and c) allow investigation into the incremental process of belief updating as predicted by the IA framework. [POSSIBLE TO INCLUDE HERE IF THIS IS INTRODUCED AS A SECONDARY OBJECTIVE WHEN DESCRIBED IN THE METHODS: In setting the three exposure conditions we also noted a fourth possible investigation, that is, to test for the presence of "shrinkage" as first discussed in (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). In implementing the study this last objective could not be satisfactorily answered therefore we leave its elaboration to the discussion section.]

In consonance with previous studies we find that listeners changed their categorization behavior in the direction of the shift in the exposure talker's VOT distributions. This provides new evidence that listeners do respond to talker statistics when the stimuli are more human-like and sampled from distributions that replicate the variability one would encounter in real life. In test block 1 participants in all groups converged on the same prior categorisation function but then their boundaries spread apart after the first exposure block. Regression analysis showed evidence in favour of the differences in boundary estimates between conditions in test blocks 2 to 4, and these differences were consistent with the direction of the distributional shift. The +10ms condition had a boundary to the right of the baseline condition and the +40ms group had a boundary right of the +10ms condition. This order of the boundary placements was maintained throughout all test blocks after the onset of exposure but their differences began to narrow from test block 5 suggesting a dissipation of distributional learning without further informative exposure.

A second finding from this study which remained opaque in previous work was that categorization differences between the groups emerged very early on after exposure. It took as few as 48 exposure trials for a clear difference to emerge between the groups. Although we do not yet know if learning was already present prior to the 48 trials, that it does not take hundreds of exposures for listeners to exhibit changes in categorizations aligns with other speech adaptation studies employing different paradigms such as perceptual recalibration and L2 accent adaptation (Bradlow and Bent (2008); Clarke and Garrett (2004); (**norris2006?**)).

We found some evidence for incremental change in categorisation boundaries as listeners received more input of the talker's cue distributions although this was not always clear from one block to another due to the uncertainty in boundary estimates. Looking at the PSE estimates at each block as a proportion of the ideal boundary implied by their respective distributions (labels Fig. 6), in the +40ms condition listeners increased the shift by roughly 10 percent in the third test block (after 96 exposure trials) from the second block but appeared to regress slightly in test block 4. In the +10ms condition boundaries did shift incrementally after each exposure block buthe proportion of while in the baseline condition, listeners showed a slight regression in test block 3 before increasing their shift towards the implied boundary in test block 4. These mixed patterns between the conditions do not clearly tell us

In this experiment we also found that the bulk of the maximum boundary shift that each group would make by the end of all 144 exposures was achieved after the first 48 exposure trials. In the +40ms condition listeners achieved their maximum shift in test block 3

What is common to all three conditions is that none of the groups converged on the category boundary implied by the exposure distributions of their respective conditions.

To understand this pattern, it is helpful to relate our exposure conditions to the distribution of VOT in listeners' prior experience. Figure 7 shows the mean and covariance of our exposure conditions relative to the distribution of VOT by talkers of L1-US English (based on Chodroff & Wilson, 2018). This comparison offers an explanation as to why the baseline condition (and to some extent the +10 condition) shift leftwards with increasing exposure, whereas the +40 condition shifts rightwards: relative to listeners' prior experience our baseline condition actually presented lower-than-expected category means; of our three exposure conditions, only the +40 condition presented larger-than-expected category means. That is, once we take into account how our exposure conditions relate to listeners' prior experience, both the direction of changes from Test 1 to 4 *within* each exposure condition, and the direction of

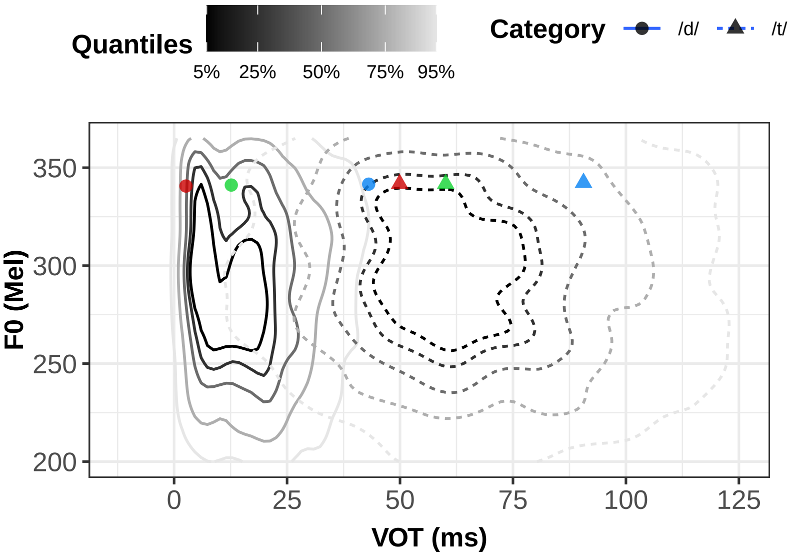


Figure 7. Placement of exposure stimuli relative to an estimate of typical phonetic distributions for 6914 word-initial /d/ and /t/ productions in L1-US English (based on 92 talkers in Chodroff & Wilson, 2018). The outermost contour of each category shows the 95% density quantile. Points show the category means of the exposure condition.

differences *between* exposure conditions receive an explanation.

Second, the reason for the slight decrease in the difference between the +10 and baseline conditions observed in Tables 1 and 2 (visible in Figure 6D as the decreasing difference between the green and red line) is *not* due to a reversal of the effects in the +10 condition. Rather, both conditions are changing in the same direction but the baseline condition stops changing after Test 2, which reduces the difference between the +10 and baseline conditions (see Table 1). The comparison across blocks thus suggests a rather uniform picture across all exposure conditions: participants' responses initially changed rapidly with exposure; with increasing exposure, these changes did not only slow down but seem to hit a hard constraint. Participants in the leftwards-shifted baseline condition did not exhibit any further changes in their categorization responses beyond Test 2. Similarly, participants in the rightwards-shifted +40 condition did not exhibit any further changes in their categorization responses beyond Test 3. Only participants in the leftward-shifted +10 condition still exhibit changes across blocks even form Test 3 to 4. But, perhaps tellingly, those participants also never reached the degree of shift that was evident in the baseline condition.

2.5.5 Constraints on cumulative changes

Finally, Figures 6C & D also compare participants' responses against those of an ideal observer that has fully learned the exposure distributions.

3 General discussion

- discuss consequences of findings for other accounts (decision-making; normalization)
- discuss fact that test stimuli deviate from exposure stimuli to different extent. on the one hand, it's just 1/4 of all trials. on the other hand, we do see relatively systematic changes in slopes each time we test. so there is evidence that even these 12 trials can affect categorisation slopes (though it is worth keeping in mind that this is a comparison across different sets of stimuli). could this explain shrinkage? unlikely since it wasn't the case in kleinschmidt and jaeger. could it explain the constraint on adaptation? that's less clear. we can, however, compare the relative mean of exposure and test. future studies could rerun the exact same paradigm but only test at position x (i.e., a between-subject version of our design)
- could some form of moving window with historical decay explain the findings? On the one hand if the moving window is very small, that would not explain why we do see some *cumulative* changes across blocks (window must be at least $48 + 12 = 60$ trials). on the other hand, the qualitative changes in the PSEs and slopes suggest that 12 trials can be enough to change some aspects of the categorisation function. it's thus *possible* that something that ways recent input much more strongly but also considers less recent input beyond 48 trials might explain the overall pattern.
- discuss potential that observed adaptation maximizes accuracy under the choice rule. use psychometric function fit during unlabeled exposure trials to calculate *accuracy* (not likelihood) on labeled trials under criterion and under proportional matching decision rules. compare against accuracy if ideal observers categorization functions are used instead.

4 References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4), 2031–2039.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, *22*, 916–943.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bejjanki, V. R., Beck, J. M., Lu, Z.-L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, *14*(5), 642–648.

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

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

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

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

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2).

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

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

Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for bayesian inference in speech perception. *The Cognitive Neurosciences*, 177–189.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in spanish and english. *Journal of Phonetics*, *49*, 77–95.

Docherty, G. J. (2011). The timing of voicing in british english obstruents. In *The timing of voicing in british english obstruents.* De Gruyter Mouton.

Fenn, K. M., & Hambrick, D. Z. (2013). What drives sleep-dependent memory consolidation: Greater gain or less loss? *Psychonomic Bulletin & Review*, *20*, 501–506.

Flege, J. E., & Brown Jr, W. S. (1982). The voicing contrast between english/p/and/b/as a function of stress and position-in-utterance. *Journal of Phonetics*, *10*(4), 335–345.

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

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

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

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

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

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.

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

Kleinschmidt, D., & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 10–19.

Kleinschmidt, D., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*.

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

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

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

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

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

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in english stops. *Language and Speech*, *10*(1), 1–28.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70.

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1562.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing*

({PhD} dissertation). The University of Iowa.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196.

Nixon, J. S., Rij, J. van, Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.

Norris, D., & McQueen, J. M. (2008). Shortlist b: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

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

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pierrehumbert, J. (2001). Stochastic phonology. *Glot International*, *5*(6), 195–207.

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

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

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

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

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

Smith, B. L. (1978). Effects of place of articulation and vowel environment on 'voiced' stop consonant production. *Glossa*, *12*, 163–175.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–136.

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112.

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

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review*, *26*, 985–992.

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

Westbury, J. R. (1979). Aspects of the temporal control of voicing in consonant clusters in english. *Texas Linguistic Forum Austin, Tex*, 1–304.

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

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064–2072.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210.

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

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

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

Zheng, Y., & Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1270.

Zue, V. W. (1976). *Acoustic characteristics of stop consonants: A controlled study.* MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB.