Listeners' categorisation behaviour correlates with gradient changes in exposure statistics

Maryann Tan[1, 2], Rachel Sabatello[2], Iva Savic[2], & T. Florian Jaeger[2]
[1]Centre for Research on Bilingualism, Stockholm University, Sweden
[2]Brain & Cognitive Sciences, University of Rochester, US

https://doi.org/10.36505/ExLing-2022/13

Abstract

Listeners can understand talkers despite cross-talker variability in the mapping from phonetic cues to linguistic categories. The mechanisms that underlie this adaptive ability are not well understood. We test to what extent listeners can adapt their interpretation of speech based on the distribution of phonetic cues in the recent input, and whether prior expectations about how talkers typically sound guide and constrain this process.

Keywords: speech perception, speech adaptation, distributional learning

Introduction

Spoken language is highly variable – a given talker's /d/ in "din" may sound more like another's /t/ in "tin". These differences arise from multiple sources including physiological, linguistic, and extra-linguistic factors. Despite the absence of invariant acoustic cues to a talker's intended category, listeners usually comprehend talkers with apparent ease. When faced with unfamiliar accents, listeners often adapt with little exposure (e.g., Bradlow & Bent, 2008).

The mechanisms underlying such adaptation are not yet fully understood. One hypothesis holds that as listeners encounter a talker, they incrementally learn the statistics of that talker's input and integrate it with their prior expectations of how talkers should sound (cf. ideal adapter, Kleinschmidt & Jaeger, 2015).

We exposed US American English (AE) listeners to recordings of AE that was phonetically manipulated between participants. All listeners heard word recordings starting with /d/ or /t/ (e.g., "dill" or "till"). Recordings varied in the primary phonetic cue to the /d/-/t/ contrast (voice onset time, VOT). Between participants, an initial exposure phase shifted the VOT distributions for /d/ and /t/ by +0, +10, or +40msecs. We assessed the consequences of those shifts during subsequent test phases that were identical across all participants.

Methods

Our approach closely follows previous work (Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; Theodore & Monto 2019) but extends these paradigms in ways intended to increase the ecological validity of the stimuli and exposure distributions. Our design choices were also motivated by intentions to computationally model the incremental changes in listener behaviour at each phase of exposure. Here, however, we report empirical observations.

Materials

Previous work employed stimuli that sounded robotic (Clayards et al.; Kleinschmidt & Jaeger, 2016) and/or exhibited unnatural cue correlations (Theodore & Monto, 2019). We used a Praat script (Winn, 2020) to create three human-sounding minimal pair VOT continua (dill-till, dip-tip, and din-tin) from original voice recordings of a 23-year-old female AE native speaker. The continua ranged from -100ms to +130ms VOT in 5ms steps. To avoid unnatural correlations with secondary cues to onset stop voicing in AE, we set the F0 at vowel onset to follow its natural correlation with VOT, as observed in the original recordings. Similarly, the duration of the vowel was set to follow the natural trade-off relation with VOT reported in Allen & Miller (1999).

Design

To assess incremental changes in listeners' categorization functions, we employed a multi-block exposure-test design (Fig. 1). Exposure was manipulated between participants. We first estimated listeners' expectations for a typical talker's VOT means and variances of /d/ and /t/. These estimates were based both on a norming experiment with our stimuli (N=24) and a phonetic database of AE onset stop voicing (Chodroff & Wilson, 2017). We then made three exposure conditions that shifted the VOT distributions for /d/ and /t/ by +0ms, +10ms, or +40ms relative to our 'typical talker' estimate (Fig. 2). Previous work set exposure distributions in the voiced and voiceless categories to have equal variance and to be distributed symmetrically around the category mean. Neither is the case for natural language. We thus sampled stimuli (see Fig. 2) from distributions with *unequal* variances observed in natural language (e.g., Chodroff & Wilson, 2017).

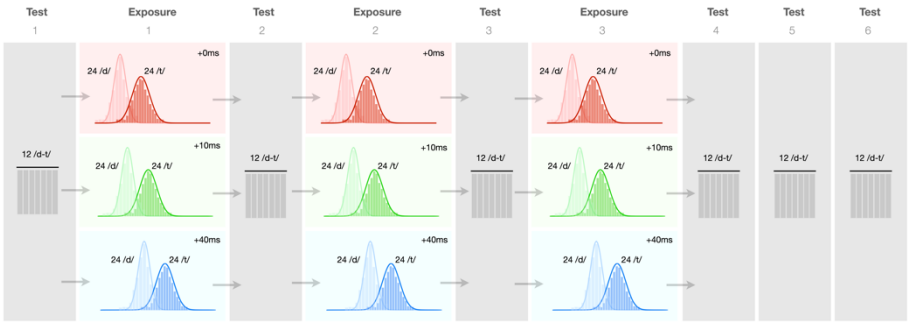

Figure 1. Block design of experiment with number of critical trials in each block. Test blocks were identical within and between conditions. Lines show the underlying distributions of the exposure and test stimuli. Block transitions were concealed from participants.

Test blocks consisted of 12 VOT-items (from -5ms–70ms), counter-balanced by minimal-pairs. In total there were 234 trials (including 18 catch trials that served to assess participant attention).

Participants

122 AE listeners (male = 61; mean age = 37.6 years, SD = 12 years) were recruited from the Prolific crowdsourcing platform, and randomly assigned to one of the three exposure conditions (+0ms, +10ms, and +40ms shift).

Procedure

Participants first underwent a headphones test and were given instructions to answer as quickly and as accurately as possible before the experiment began. On each exposure trial, participants clicked on a green button to play the recording of the word. Simultaneously, written forms of possible responses were displayed on the top left and top right of the screen. As shown in Fig. 2, half of the exposure trials *labelled* the voicing category: e.g., if a recording was intended to be *voiced*, both displayed words started with "d" (e.g., displaying "dill" and "dip" for a "d/till" recording). The other half of the trials were *unlabelled* (e.g., displaying "dill" and "till" for a "d/till" recording). Upon clicking on the word heard the next trial commenced. Test trials were always unlabelled. The order of trials and the left-right placement of responses was randomized for each participant, and counter-balanced across participants.

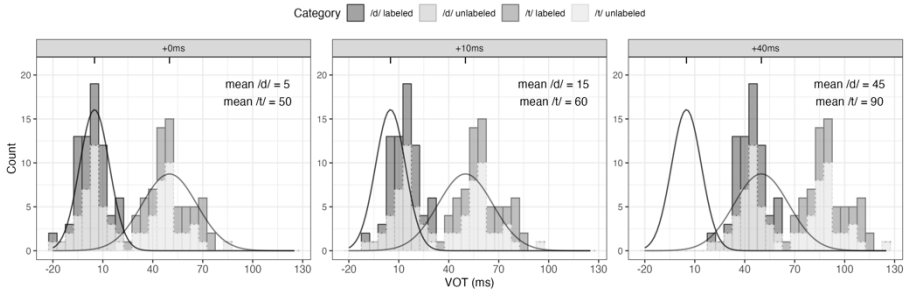

Figure 2. Histograms of distributions of /d/ and /t/ for the three exposure conditions. Black density lines are identical across panels and show the VOT distribution that the +0ms condition is sampled from.

Results and discussion

Figure 3 shows listener categorisation behaviour at each test block. We focus on the estimated category boundary (point of subjective equality, PSE). Block 1 indicates participants' PSE *before* informative exposure, indicating participants' *prior* expectations about the VOT distribution of /d/ and /t/ for this talker. Our previous estimate of this PSE (24.5, which determined the +0ms exposure distributions) proved to be about 20 ms lower than the PSE observed for Block 1 (44.7 ms).

By Block 2, the PSE of all exposure conditions had shifted. The direction and magnitude of these shifts qualitatively follow the predictions of an ideal adaptor (or similar theories of incremental adaptation). Specifically, (1) the PSEs for the three conditions order in the same way as the means of exposure distributions (+0ms < +10ms < +40ms); (2) the PSE of Block 2 shifted leftwards relative to Block 1 for the +0ms and +10ms conditions, in line with the observation that the *prior* PSE was actually about 20ms to the right of what we intended to be the +0ms condition (so that +0ms is actually -20ms exposure and +10ms is actually -10ms exposure); and, finally, (3) the PSE of Block 2 shifted rightwards relative to Block 1 for the +40ms conditions (which is actually +20ms exposure once the correct prior PSE is considered).

These PSEs largely remained unchanged through Blocks 3-6: the remaining 96 exposure trials had only minor effects that showed mostly in the most extreme shift (+40ms exposure). This suggests that participants learned the distributions of the talker quickly–after exposure to 48 trials (2nd panel). Of note is that, despite these rapid changes in PSEs, the *extent* to which PSEs changed was greatly limited: even though the +0ms and +40ms exposure distributions differed by 40ms, the PSEs for those two conditions only differed by 10-14ms.

While these results are broadly consistent with exemplar and Bayesian theories of incremental adaptation, they also raise novel questions. In particular, it is unclear why shifts are both so quick—which would seem to imply weak weighting of prior expectations—and yet strongly constrained—which would seem to imply strong weighting of prior expectations. To the best of our knowledge, this tension has not previously been discussed but strikes as an important characteristic of speech adaptation to be understood.

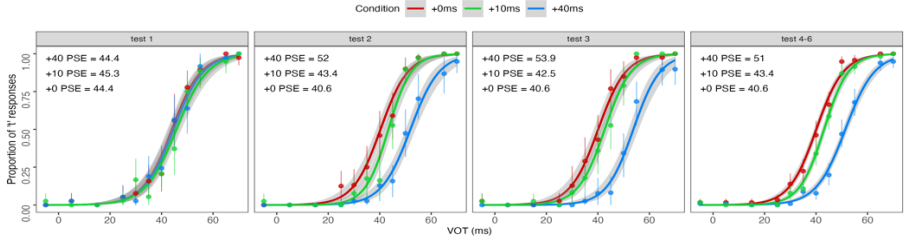

Figure 3. Mean categorisation functions by exposure condition. The last panel combines the final three post-exposure test blocks into one.

References

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).

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

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review, 26(3)*, 985-992.