Phonetic cue distributions guide perceptual adaptation in speech: Evidence from a three-week study with a natural non-native accent

Xin Xie
xxie14@uci.edu
Department of Language Science,
3151 Social Science Plaza
University of California, Irvine

Chigusa Kurumada
ckuruma2@ur.rochester.edu
Department of Brain and Cognitive Sciences,
Meliora Hall
University of Rochester

Abstract

Human languages vary widely in the combination and balance of phonetic cues used to encode speech sounds, a source of non-native accents in second language learners. It has been hypothesized that repeated exposure to an accent leads to adaptive perceptual changes in native listeners through multidimensional distributional learning. Although this hypothesis is highly influential, it has rarely been tested against the complexity of naturally produced accented speech, and for a period longer than a single-session experiment. The current large-scale (N = 338), five-session experiment goes beyond this status-quo to examine the adaptation of native American English listeners to naturally produced Mandarin-accented English. A repeated exposure-test design was used to characterize adaptive changes in perception from the first few minutes to over the course of three weeks. The results reveal that behavioral changes can be predicted by listeners' sensitivity to changes in phonetic cue distributions. Possible joint contributions of early-stage auditory normalization and later-stage decision processes are discussed.

Keywords: adaptive speech perception, non-native accents, distributional learning, repeated exposure-test paradigm, longitudinal perceptual testing

Introduction

Human perception is finely tuned to the environment and its changes. In fact, the ability to adaptively map sensory cues to visual and auditory experiences is essential for the subjective stability of perception and motion (Gibson, 1954). In speech, this adaptivity allows us to recognize the input from a wide range of individual talkers as well as those from different linguistic and social backgrounds, including nonnative accented speech (Norris, McQueen, & Cutler, 2003; Magnuson & Nusbaum, 2007; Johnson & Sjerps, 2021). For example, while speech from an unfamiliar non-native (L2) talker may initially sound confusing or difficult to understand, such difficulties can dissipate, sometimes after only a few minutes of exposure (Clarke & Garrett, 2004).

Over the past two decades, theories have hypothesized that the mechanism supporting this adaptivity relies on **distribution of underlying acoustic and phonetic cues**. Listeners and their perceptual judgments appear to be sensitive to the cue distributions over latent perceptual dimensions, unique to an individual and to an accent (Maye, Aslin, & Tanenhaus, 2008; Clayards, Tanenhaus, Aslin, & Jacobs, 2008). And models that learn to adapt internal cue-category mappings to the structure of the current input are shown to approximate behavioral changes in speech perception (Kleinschmidt & Jaeger, 2015; Theodore & Monto, 2019).

Consider, for example, the distinction between the minimal pair, ***t****ime* and ***d****ime* in English. In L1 speakers' productions, the two word-initial consonants /d/ and /t/ can largerly be distinguished on the basis of a few cue dimensions (e.g., voice onset time (VOT)), with longer VOTs for /t/ than /d/. In L2 speakers' productions, however, the means and variances of these distributions may differ from those in L1 speech, or may be compensated for by another cue (Schertz, Cho, Lotto, & Warner, 2015; Bent & Baese-Berk, 2021). Learning such unique distributional features should therefore improve the recognition of the contrast in L2-accented speech. Note that the term *distributional learning* can be used for a wide range of phenomena and paradigms, including unsupervised category induction in acquisition; Here we use the term more broadly to refer to listeners' learning of the distribution of acoustic cues for a particular sound category. It can occur in a strictly bottom-up manner (Clayards et al., 2008) or with the aid of lexical feedback (or other types of supervision), as we implement in the experiment reported below.

Although well accepted as a candidate mechanism for adaptive speech perception, distributional learning still lacks empirical tests against natural speech variability. Much of the existing support comes from resynthesized speech stimuli, typically crafted by manipulating a single cue dimension (e.g., VOT for /t/ vs. /d/). Compared to the natural input, the distribution that listeners must learn is made tractable (i.e., less complex). At the same time, the manipulation often disrupts the covariation between multiple cues—a critical component of the speech signal that is essential for human listeners (Idemaru & Holt, 2011). In general, it is still largely unknown whether distributional learning, as has been studied with parameterized and resynthesized stimuli, can scale to cross-talker or accent variations *in the wild* (Zheng & Samuel, 2020).

In addition, past research has rarely extended beyond a single experimental session. Even when multiple training sessions (Bradlow & Bent, 2008) or delayed testing (Eisner & McQueen, 2006; Zheng & Samuel, 2023) were used, adaptive changes were typically evaluated only once or twice with no additional exposure in between. However, some changes in perception are gradual in nature, occurring over time with memory consolidation in between (Xie, Earle, & Myers, 2018). Indeed, many real-life examples of regional or non-native accent accommodation emerge through multi-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

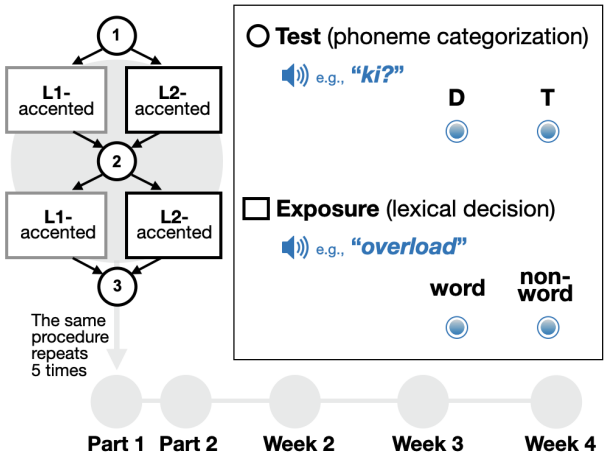

Figure 1: In each session, three brief test blocks (five items each) alternate with two exposure blocks (30 items each). During test, participants provide phoneme categorization judgments (e.g., Did you hear *kid* or *kit*?). During exposure, participants perform lexical decision task to provide word vs. non-word judgments. The same procedure repeats five times over three weeks.

ple encounters with the same talker and/or accent (e.g., routine interactions with international colleagues). It is therefore critical to track adaptive changes in perception for a broader time scale and testing schedule in order to understand whether distributional learning, or any other mechanism, can support flexible adaptation in real-world communication.

To achieve this goal, we conducted a perceptual experiment on native listeners' adaptation to an L2-accented speech across multiple episodes of encounter. We used a repeated exposure-test paradigm (Vroomen, van Linden, de Gelder, & Bertelson, 2007) over five sessions spanning three weeks. In each session, participants completed three short test blocks alternating with two exposure blocks (Fig.1). Half of them were exposed to L1-accented English [control] and the other half to L2-accented English [target], and all were tested on novel L2-accented stimuli. All items were naturally produced tokens of L1-/L2-accented speech, with which we investigated changes of categorization judgments as the input accumulates within and across sessions.

A case in point: Syllable final voicing contrast in Mandarin-accented English

Distributional learning theories predict that the acoustic-phonetic properties of categories experienced during exposure and test have direct influence on *whether* and *how* listeners adapt. Xie and Kurumada (2023) sought to test this prediction by focusing on a specific accent feature with well-characterized acoustic cue distribution, syllable-final voicing contrast (Eisner, Melinger, & Weber, 2013). To illustrate this, Fig.2A shows a Mandarin-accented /d/- vs. /t/-final words (points) superimposed on distributions (ellipses) simulated from a corpus of L1-accented English (Xie, Theodore, & Myers, 2017). As shown by the two ellipses, /d/s in L1-accented speech often have an overall shorter closure and longer vowel duration than those for /t/. However, many instances of the L2 accented /d/ tokens fall within the /t/ distribution or where it overlaps with the /d/ distribution. In consequence, an accented syllable-final /d/ (e.g., *kid*) often sounds like a /t/ (e.g., *kit*) to L1 English listeners unfamiliar with the accent. Critically, a third cue (burst duration) with little salience in L1-accented speech (Fig.2B) becomes the primary cue in Mandarin-accented English and is diagnostic of the contrast (Fig.2C). If adaptive changes in perception involve distributional learning, then categorization of this target contrast should improve as native listeners gradually up-weight the burst dimension over the other cues.

Methodologically, probing the effects of distributional learning can be difficult, especially with repeated exposures and tests. Most critically, repeated encounters with relevant accent features in the *test* items, even without feedback, serve as additional sources of accent exposure. That is, repeated or prolonged testing itself may dilute or cancel out the effect of the exposure conditions. Each test block must therefore be kept short. However, this may compromise the statistical power of the results. Additionally, Xie and Kurumada (2024) found that categorizations of L2-accented test items improved even in the L1-accent exposure (i.e., control) condition, likely due to the fact that a small number of minimal pairs presented during test provided perceptual anchors, which scaffold listeners' categorization judgments.

To address this conundrum, our current test stimuli were limited to as few as five minimal pairs and only one member of each pair was presented to a given participant. Unlike in a canonical approach where exposure and test items are selected based on perceived levels of accentedness, we adopted a simulation-based approach to select five minimal pairs that are likely to provide an effective test of the distributional learning hypothesis. We trained Bayesian ideal observer models on the L1- vs. L2-accented exposure items annotated for the three cues: vowel, closure, burst (Tan, Xie, & Jaeger, 2021). These models were then used to predict human categorization judgments on all the 60 minimal pair test items from Xie et al. (2017). This allowed us to rank the pairs in terms of the relative advantage expected from the distributional information provided in the L2-accented exposure but not in the L1-accented exposure. In other words, we simulated the theoretical maxima of distributional learning for each pair in the two exposure conditions and selected the five pairs that maximized the likelihood of detecting the difference between conditions. For details on the simulation and stimulus selection, see Tan et al. (2021) and Xie and Kurumada (2024).

Experiment

To mitigate the risk of participant attrition, we conducted the current experiment online, allowing participants to complete

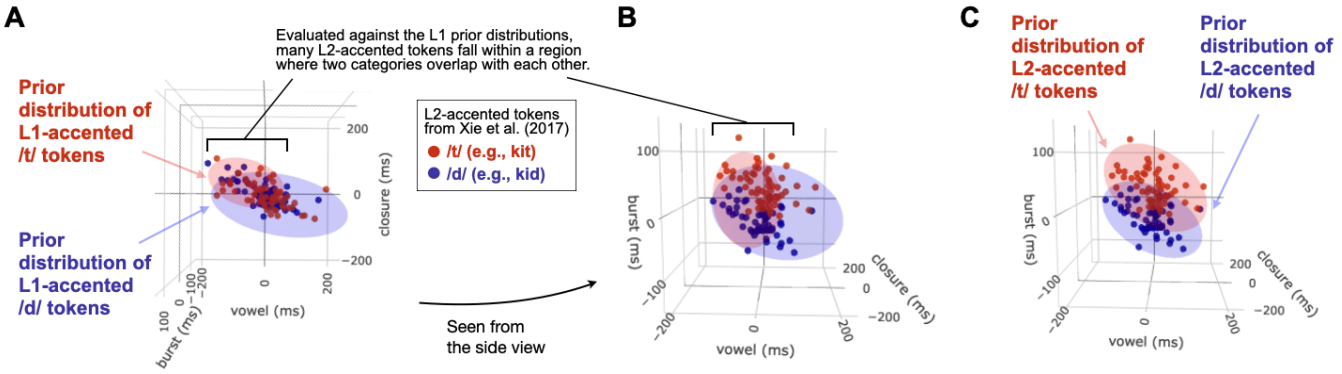

Figure 2: A: L2-accented talkers' productions (points) plotted against L1-accented /d/ and /t/ categories in a two-dimensional talker-normalized phonetic space (vowel and closure duration; for details, see Xie et al., 2017 and Tan et al., 2021). Each point correspond to one L2-accented token (blue = /d/, red = /t/). The ellipses represent the L1-accented /d/ and /t/ categories, showing 95% probability density of multivariate Gaussian categories. Many L2-accented tokens fall within an 'ambiguous' region of the acoustic-phonetic space where L1-accented /d/ and /t/ overlapped. B: Same as Panel A, but seen from a side view with the third dimension, burst duration. C: Same as B except that the /d/ and /t/ categories depict the distribution of L2-accented productions. Learning this distribution is expected to lead to more accurate categorization.

the experiment from the comfort of their homes and providing flexibility in scheduling. The number of participants was determined based on the effect size and attrition rate from a preliminary study (Xie & Kurumada, 2024). The resulting sample size is one order of magnitude larger than what is typically used in similar adaptation studies.

Participants

338 monolingual, native speakers of American English, aged 18-45, were recruited via Prolific (https://www.prolific.co/) and completed Session 1 of the experiment via the online testing platform FindingFive (https://www.findingfive.com/). They were randomly assigned to the L2-accented exposure condition ([target] n= 160) or the L1-accented exposure condition ([control] n = 178). Of those, 214 participants (63.3 %) completed all five sessions (n = 105 in the L2-accented exposure condition; n =109 in the L1-accented exposure condition). The attrition rate over three weeks was comparable between the two exposure conditions (35% and 39%, respectively). 5% of the participants (12 out of 214) reported that they regularly hear Mandarin Chinese spoken by a family member or a close friend. Those subjects were excluded from the analysis. The final dataset included 202 participants (98 and 104 in the L2- vs. L1-accented exposure conditions, respectively) who completed all five sessions.

Stimuli

All stimuli were taken from Xie et al. (2017). Exposure stimuli used for the L2-accented exposure condition consisted of 90 English words (30 critical and 60 filler items) and 90 phonotactically-legal nonwords. The critical items were all multisyllabic words ending in /d/ (e.g., *overloa**d***, *lemona**de***). The fillers and non-words did not contain any /d/ or /t/ sounds, and no stop sounds other than /d/ appeared in the final position of the word. The exposure list for the L1-accented exposure condition was identical except that the /d/-final words were substituted with non /d/-final filler words, and all items were produced by an age- and gender-matched native speaker of American English. The 180 exposure items were evenly distributed across the two exposure blocks. Word-block assignment was counterbalanced across participants and remained constant within participants across the five sessions. Item presentation was randomized within each block.

The test stimuli consisted of five pairs of monosyllabic English words that differed in the final stop consonant (e.g., kid-kit). These 10 stimuli (five pairs of /d/ vs. /t/) were organized into eight lists of five stimuli each, such that only one member of each minimal pair was presented to each participant. This was done to prevent listeners from anchoring their judgments solely on the minimal pair comparisons within a test block. And the odd number of test items was expected to discourage them from choosing /d/ vs. /t/ an equal number of times in each block. Each list contained either three /d/ and two /t/ items (e.g., *feed, kid, wed*, *fright, plot*) or two /d/ and three /t/ items (e.g., *feed, plod*, *fright, kit, wet*).

Procedure

Five experimental sessions were administered on five days over the course of three weeks (Fig.1). Each session consisted of three test blocks (five trials each) interleaved with two exposure blocks (90 trials each). After a headphone check to adjust volume and confirm the audibility of the audio stimuli, participants began with a test block. Participants were informed that during this block they would hear words ending in /d/ or /t/ and asked to provide two-alternative forced choice (2AFC) responses to the question "Did you hear a D

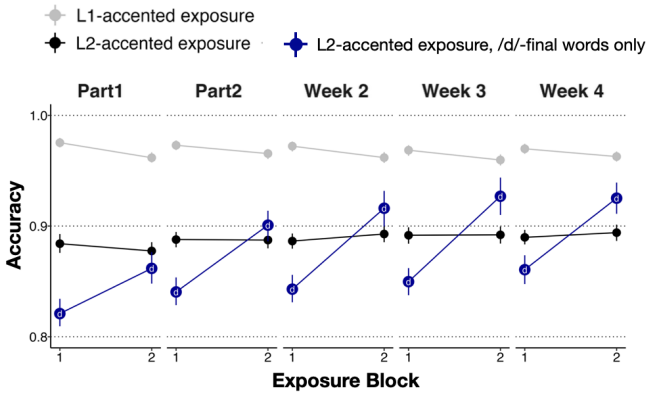

Figure 3: Recognition accuracy during exposure. Blue points represent the accuracy for the /d/-final (critical) words (e.g., *overload*). Error bars represent bootstrapped 95% confidence intervals over by-participant means.

or a T?" One of the eight lists, which all contain five items, was presented in each test block. List assignment was counterbalanced across participants and kept constant within and across sessions within participants.

During exposure, participants completed a lexical decision task (i.e., word or nonword). Participants heard one token at a time and responded whether it was a real word of English (e.g., *lemonade*) or a nonword (e.g., *salvary*). In the first session (Day 1), after the five test/exposure blocks, all participants completed a questionnaire about their language background and familiarity with L2 accents. Participants were subsequently invited back to the experiment four more times. In each session, the experiment was open for 48 hours, starting at 9 a.m. Pacific time on a given day. The interval between the first and second sessions was 24-48 hours and one week thereafter, with the exact interval length varying between participants. Participants who missed a session were removed from the experiment.

Results

Exposure Accuracy

Fig.3 shows accuracy on the lexical decision task during exposure. As expected, performance in the L1-accented exposure condition was consistently near the ceiling (1st block: mean = .95, SD = .07, last block: mean = .95, SD = .05). In contrast, the task was more difficult for participants in the L2-accented exposure condition (1st block: mean = .85, SD = .05, last block: mean = .86, SD = .06). Notably, recognition of the /d/ final words increased steadily within each session and across sessions (1st block: mean = .73, SD = .09; last block: mean = .90, SD = .10), despite the generally increased task difficulty in the L2-accented exposure condition compared to the L1-accented exposure condition. The apparent drop in accuracy between sessions is largely due to the fixed item-block assignment, with block 2 items being easier than block 1 items. Accuracy improvement is evident when tracked separately for a given block across sessions. Overall, the incremental improvement in the L2-accented exposure group replicates Xie and Kurumada (2024), suggesting that (1) even 15 critical items per exposure block were sufficient to improve recognition of the L2-accented feature, and (2) these improvements accumulated with increasing exposure.

Test Accuracy

As predicted, recognition of /d/-final words (e.g., *kid*) was initially much less accurate than recognition of /t/-final words (e.g., *kit*) (Fig.4). Also as predicted, recognition accuracy for /d/-final words increased steadily from .29 (SD = .28) on Day 1 to .55 (SD = .33) on the last day of Week 4 in the L2-accented exposure group (+ 1.10 log-odds). A similar improvement was seen in the L1-accented exposure group, albeit to a lesser extent (+ .60 log-odds), from .30 (SD = .30) to .44 (SD = .34).

We fit a mixed-effect logistic regression to the test data using the lme4 package in R. The analysis predicted accuracy on a test item (1 = correct, 0 = incorrect) from the full factorial of exposure condition (effect-coded, -.5 = L1-accented exposure vs. +.5 = L2-accented exposure), category (effect-coded, -.5 = /t/- vs. +.5 = /d/-final words), and test block (1-15 as a numeric variable, scaled by dividing by two standard deviations). Coding test block as a numeric variable allowed us to examine whether incremental, repeated exposure resulted in cumulative improvement in the test performance. We began with the maximal random effect structure justified by the design and stepwise removed higher-order interactions in the event of convergence failure. The final model included random by-participant intercepts and slopes for category, as well as by-item intercepts and slopes for exposure condition, category, and their interaction.

The overall accuracy of both groups improved significantly over time, as indicated by a significant main effect of (test) block ($\hat{\beta}$ = .31, SE = .05, z = 6.08, p $<$ .002). Additionally, the block by category interaction was highly significant ($\hat{\beta}$ = 1.16, SE = .10, z = 11.4, p $<$ 2e-16), suggesting that the magnitudes of improvement were different betwen the /t/-final and /d/-final words. A follow-up simple effects analysis showed that recognition accuracy increased significantly for /d/-final words ($\hat{\beta}$ = .89, SE = .07, z = 12.8, p $<$ 2e-16) and decreased significantly for /t/-final words ($\hat{\beta}$ = -.27, SE = .07, z = -3.64, p $<$ .001). Thus, the overall improvement across blocks was driven by the greater improvements for /d/-final words than the loss of accuracy for /t/-final words.

Most importantly, the same simple effects analysis revealed that the condition by block interaction was significant for both /d/- and /t/-final words (/d/: $\hat{\beta}$= .39, SE = .14, z = 2.86, p $<$ .005; /t/: $\hat{\beta}$ = .45, SE = .15, z = 3.07, p $<$ .003). That is, exposure to the target L2-accent compared to an L1-accent resulted in more accurate recognition of the /d/ vs. /t/ contrast overall.

Cue-based analyses

To investigate whether repeated accent exposure increases listeners' reliance on L2 accent-specific cue distributions, we predicted the log-odds of a /d/ over a /t/ response with three durational values of a given test item: vowel, closure, and burst (Fig.2). If exposure leads to enhanced sensitivity to the underlying distributional properties, burst duration should be increasingly more predictive of listeners' /d/ responses in the L2-accented exposure condition over time, compared to the L1-accented exposure condition.

We fit a mixed-effect logistic regression similar to the one reported above. Instead of accuracy, this model predicted /d/-responses (1 = /d/, 0 = /t/) from the standardized acoustic cue values of vowel, closure, and burst, as well as exposure condition (effect-coded, -.5 = L1-accented exposure vs. +.5 = L2-accented exposure), and test block (1-15, as numeric). The model included random by-participant intercepts and slopes for test block, as well as by-item intercepts and slopes for exposure condition, test block, and their interaction.

The main effects of the three acoustic cues are all highly significant (vowel: $\hat{\beta}$ = .78, SE = .09, z = 8.95, p $<$ 2e-16; closure: $\hat{\beta}$ = 1.61, SE = .15, z = 10.94, p $<$ 2e-16, burst: $\hat{\beta}$ = - 1.87, SE = .09, z = -20.82, $<$ 2e-16), as well as their interactions with the test block (vowel*block: $\hat{\beta}$ = .29, SE = .14, z = 2.06, p $<$ .04; closure*block: $\hat{\beta}$ = .83, SE = .24, z = 3.53, p $<$ .001, burst*block: $\hat{\beta}$ = - .64, SE = .15, z = -4.31, p $<$ .0001). The results suggest that items with longer vowel and closure durations and shorter burst durations are more likely to elicit /d/ responses, and this tendency increased over time. Most critically, the model suggested a significant three-way interaction between burst * condition * block ($\hat{\beta}$ = -.71, SE = .27, z = -2.64, p $<$ .009). Follow-up simple-effects analyses confirmed that only listeners with the L2-accented exposure ($\hat{\beta}$ = -.45, SE = .16, z = -2.88, p $<$ .004), not those in the L1-accented exposure condition (p $>$ .30), increased their reliance on burst as a cue to categorization over time. No between-condition differences were observed for vowel and closure durations, suggesting that the two conditions did not differ in terms of how these two cues affect categorization.

General Discussion

Due to the unique cue distributions, the syllable-final /d/-/t/ contrast in Mandarin-accented English can be initially ambiguous for native listeners, as seen here: While /t/ items were correctly recognized, /d/ items were often misheard as /t/. However, repeated exposure to naturally produced accented speech led to better recognition of these initially ambiguous sounds. We observed changes in recognition accuracy with greater granularity and over a longer time period than any previous study. The changes began after the first few minutes of exposure and continued over five sessions. They were *rapid* (starting after the first exposure block), *sustained* (no significant decline after a delay), and *monotonic* (stable increase of accuracy within and across blocks). Taken together, the results offer clear evidence that continued exposure to a non-native accent results in cumulative improvements.

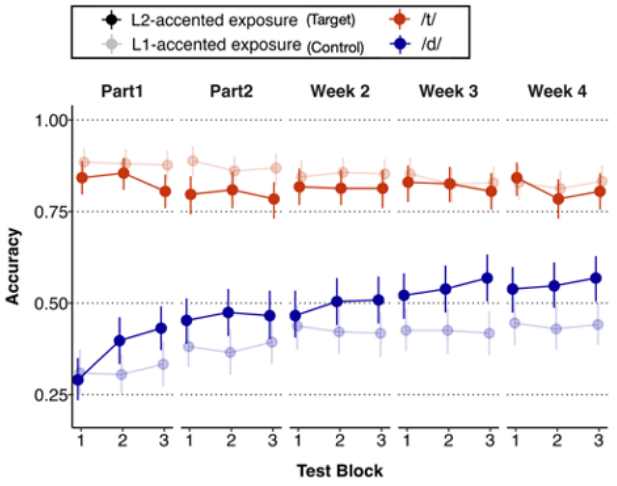

Figure 4: Recognition accuracy during test, aggregated over items and participants. Error bars represent bootstrapped 95% confidence intervals over by-participant means.

Most importantly, the phonetic cue-based analysis lends credence to the idea that the adaptive perceptual changes in the L2-accented exposure condition reflect listeners' ability to track the underlying cue distributions. Evaluated against listeners' long-term experiences, most of the accented test tokens initially fall within the /t/ category. However, repeated exposure resulted in increased perceptual weights on the burst dimension. The fact that the burst duration of the test items became increasingly more predictive of /d/-responses in the L2-accented exposure condition than in the L1-accented exposure condition supports the conclusion that the changes in listeners' categorization reflect an expected result of distributional learning.

The current results extend the existing knowledge in several important ways. Chief among them is that distributional learning can withstand the perceptual complexity of non-native accent, a premise that has previously been questioned (Zheng & Samuel, 2020). Earlier seminal work using the dimension-based statistical learning paradigm (Idemaru & Holt, 2011) showed that listeners can rapidly down-weight secondary phonetic cues of a contrast (e.g., f0 for stop voicing) when they conflict with expectations from primary cues (e.g., voice onset time) or long-term lexical knowledge. Such re-weighting can be detected after as few as eight trials of exposure to an altered distribution (Hodson, Shinn-Cunningham, & Holt, 2023). The current finding extends this in an ecological test: listeners can not only down-weight familiar cues, but also up-weight a cue dimension that is not expected to be as informative as the familiar cues based on prior experience. And the learning appears to begin in the first block, i.e., after $\sim$15 instances of accent-specific cue exposure. The remarkable speed merits further exploration of underlying computational and memory encoding mecha-

nisms that support multidimensional cue learning (Goudbeek, Swingley, & Smits, 2009; Ashby & Maddox, 2011).

The rapidity is even more remarkable when combined with the longevity of learning effects. The five sessions of the experiment were separated by up to six days in between, likely with a substantial amount of intervening input of L1-accented (and perhaps other L2-accented) speech. A recent study using a delayed test of perceptual recalibration to a fricative contrast also reported a week-long retention, albeit much weaker (Zheng & Samuel, 2023). If distributional learning assumed here was agnostic to the identity of the talker or talker group, each new token of perceptual input would continually overwrite the previously learned distribution, interfering with retention. Instead, listeners seem to be able to represent each unique distribution for a talker or talker group and retrieve it as needed. Our study employs natural L2 accents, making it easy for listeners to detect the nonnativeness of the to-be-learned talker. This kind of social labeling might have helped listeners to separate experiences with this talker from experiences with other L1-accented talkers in their daily communication. This is consistent with the idea that listeners construct multiple internal models (e.g., for L1-accented vs. L2-accented speech) and switch between them as a way to avoid cross-talker or cross-accent interference (Xie, Liu, & Jaeger, 2021). Elucidating the nature of such models and possible ways in which learning generalizes across models will be a productive avenue for theory building.

Several other questions remain unanswered. First, we observed that test response accuracy in the L1-accented exposure condition showed significant improvements over time. But how? One possibility is that distributional learning also occurred in the L1-accented participants, but only from exposure to the test items. Alternatively (but not mutually exclusive), listeners may simply increased their /d/ responses over time to approximately 50-50 responses between /d/ and /t/. Given that their initial responses were heavily biased toward /t/, any form of response equilibration that approaches 50% /d/ responses could predict the *improvements*. To address this, our future iteration of the experiment will remove any repeated use of test items and could vary the number of /d/ and /t/ items across test blocks.

Second, and related, improvements in recognition accuracy across test blocks may be supported by more than one mechanism. Although the current results strongly favor the involvement of distributional learning in speech adaptation, this does not negate that listeners engage multiple mechanisms to aid their categorization. Changes in decision biases mentioned above are one such way. It is also possible that auditory perception adjusts to a talker's baseline at the prelinguistic level, as postulated in theories of talker-based normalization (McMurray & Jongman, 2011; Zhang & Chen, 2016; Choi, Hu, & Perrachione, 2018), contributed to the observed behavioral changes over time. This may occur simultaneously with, but independently of, distributional learning. Existing experimental results, both behavioral and neuroimaging, are limited in the ability to tease apart these different possibilities (Xie, Jaeger, & Kurumada, 2023). A novel approach to experimental design is needed to tease out the relative (and joint) contributions of these multiple mechanisms. The longitudinal data collected in the current study enables more effective model comparison compared to relying on a one-time snapshot of human behavior. This is achieved by reducing overfitting to a specific time point, enabling better cross-validation across time points, and, most importantly, allowing for more accurate parameterization.

Finally, it is also important to recognize that the longer-term changes in perception may involve different memory and learning mechanisms than those optimal for short-term adaptation. In the real world, repeated exposure to an accented talker often provides not only the distributional information about a particular contrast (e.g., syllable-final voicing), but also a multitude of other features that co-vary with this contrast (e.g., other aspects of phonological accent, lexical choice, common grammatical errors, and visual or situational contexts of exposure). The *relearning* of acoustic distributions may simultaneously strengthen its links with these co-varying features, gradually forming a long-term, holistic memory representation of a given accent. If this is the case, what it means to "adapt" to an accent may change qualitatively over time: Initially, listeners need to learn different bottom-up distributional structures. After some exposure, they simply need to select and switch between these internal representations (e.g., one for a Mandarin accent and another for a Spanish accent). Indeed, neuroscientific evidence suggests that strong activation of neural ensembles for long-term memory typically occurs after prolonged exposure. In contrast, brief and isolated exposures result in weaker activations that are more susceptible to memory pruning (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014). The reactivation and maintenance of selective memory traces during relearning (Ryan & Frankland, 2022) may be crucial to help listeners create multiple internal representations and select between them as they navigate variability in speech input. Further longitudinal experiments are needed to test this possibility.

In summary, new evidence supports the hypothesis that perceptual flexibility exploits sensitivity to input statistics. However, it is still unclear whether distributional learning is the sole mechanism or whether it works in concert with other mechanisms. Future studies should thus consider a wider range of accent features and time scales to provide a more complete picture. The current phonetically informed, simulation-based approach is scalable and will provide a principled way forward to identify the mechanisms underlying adaptive speech perception.

References

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161.

Bent, T., & Baese-Berk, M. M. (2021). Perceptual learning

of accented speech. *The Handbook of Speech Perception*, 428-464. doi: 10.1002/9781119184096.ch16

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

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics*, *80*, 784–797.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*, 3647–3658. doi: 10.1121/1.1815131

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804-809. doi: 10.1016/j.cognition.2008.04.004

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*, 1950-1953. doi: 10.1121/1.2178721

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*, 148. doi: 10.3389/fpsyg.2013.00148

Gibson, J. J. (1954). The visual perception of objective motion and subjective movement. *Psychological Review*, *61*, 304-314. doi: 10.1037/h0061885

Goudbeek, M., Swingley, D., & Smits, R. (2009, 12). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of experimental psychology. Human perception and performance*, *35*, 1913-1933. doi: 10.1037/a0015781

Hodson, A. J., Shinn-Cunningham, B. G., & Holt, L. L. (2023, 9). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. *Cognition*, *238*, 105473. doi: 10.1016/j.cognition.2023.105473

Idemaru, K., & Holt, L. L. (2011, 12). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1939-1956. doi: 10.1037/a0025641

Johnson, K., & Sjerps, M. J. (2021). Speaker normalization in speech perception. In *The handbook of Speech Perception* (p. 145-176). John Wiley & Sons, Ltd. doi: https://doi.org/10.1002/9781119184096.ch6

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014, 6). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*, 8997-9002. doi: 10.1073/pnas.1319438111

Kleinschmidt, D. F., & Jaeger, T. F. (2015, 4). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148-203. doi: 10.1037/a0038695

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 391-409.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science*, *32*, 543-562. doi: 10.1080/03640210802035357

McMurray, B., & Jongman, A. (2011, 4). What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219-246. doi: 10.1037/a0022325.What

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

Ryan, T. J., & Frankland, P. W. (2022, 3). Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, *23*, 173-186. doi: 10.1038/s41583-021-00548-3

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183-204.

Tan, M., Xie, X., & Jaeger, T. F. (2021). Using rational models to understand experiments on accent adaptation. *Frontiers in Psychology*, *12*, 1-19. doi: 10.3389/fpsyg.2021.676271

Theodore, R., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin and Review*, *26*, 985-992. doi: 10.3758/s13423-018-1551-5

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

Xie, X., Earle, F. S., & Myers, E. B. (2018, 2). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*, 196-210. doi: 10.1080/23273798.2017.1369551

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review. *Cortex*, *166*, 377-424. doi: 10.1016/j.cortex.2023.05.003

Xie, X., & Kurumada, C. (2023). Nonnative accent adaptation in the initial moments and over a month. *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*.

Xie, X., & Kurumada, C. (2024, 5). From first encounters to longitudinal exposure: a repeated exposure-test paradigm for monitoring speech adaptation. *Frontiers in Psychology*, *15*. doi: 10.3389/fpsyg.2024.1383904

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of non-native speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*, e22–e56. doi: 10.1037/xge0001039

Xie, X., Theodore, R., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 206-217. doi: 10.1037/xhp0000285

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1252.

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*, 1270-1292. doi: 10.1037/xlm0000788

Zheng, Y., & Samuel, A. G. (2023, 3). Flexibility and stability of speech sounds: The time course of lexically-driven recalibration. *Journal of Phonetics*, *97*. doi: 10.1016/j.wocn.2023.101222