NONNATIVE ACCENT ADAPTATION IN THE INITIAL MOMENTS AND OVER A MONTH

Xin Xie[1] & Chigusa Kurumada[2]

[1]Department of Language Science, University of California, Irvine;
[2]Department of Brain and Cognitive Sciences, University of Rochester
xxie14@uci.edu, ckuruma2@ur.rochrster.edu

ABSTRACT

Perceptual difficulties associated with unfamiliar talkers or accents are known to dissipate—sometimes within minutes of exposure. How do the adaptive changes in perception evolve beyond these initial moments of encounter? The current study examined incremental changes in native listeners' recognition of L2-accented speech (Mandarin-accented English) over two timescales: within the first few minutes (Exp.1) and across five sessions spanning a month (Exp.2). We developed a new repeated exposure-test paradigm to track the trajectory of recognition improvements for a category that was initially confusable with another (i.e., word-final /d/ as in *feed* sounding like /t/ as in *feet*). L1 listeners' adaptation was detected as early as after ~30 instances of exposure and continued to evolve over the subsequent sessions, suggesting a common process guiding the rapid adaptation to, and the longer-term accommodation of, an initially unfamiliar accent.

Keywords: Speech perception, nonnative (L2) accents, short- and long-term adaptation, repeated and incremental testing, computational modelling

1. INTRODUCTION

How listeners navigate the substantial amount of cross-talker variability is a central question in speech perception. The "same" phoneme or word is produced with distinct acoustic-phonetic properties across talkers with different characteristics (e.g., height, gender, accent). This variability is known to make the recognition of an unfamiliar talker or accent difficult, [1], [2]. Such difficulties can, however, dissipate as listeners adapt to the current input [3]–[6]. For example, native listeners of English become faster and more accurate in responding to Spanish- and Mandarin-accented English after initial exposure to these previously unfamiliar accents within as few as two to four sentence-length utterances [4], [7].

Alongside the short-term adaptation, real-world speech recognition and accent adaptation also evolve over episodes of social interaction across talkers and contexts [8], [9]. Long-term experiences with an accent gained over months and years have been reported to facilitate the comprehension of, and adaptation to, a novel talker from a similar accent background e.g., [10], [11]. However, environmental exposure to a previously unfamiliar accent alone does not always lead to a significant change of perception [12]. Currently, little is known about the nature and the amount of exposure needed to support the L1-/L2-accent accommodation. Questions also remain open about whether rapid adaptation seen in the initial moments of encounter and the longer-term accent accommodation draw on a single or multiple distinct mechanism(s) [13].

This is the knowledge gap the current study set out to address. Specifically, we asked two questions: (Q1) ***How rapid is the adaptation?*** In the initial moments of encountering an *a priori* unfamiliar accent, how do the recognition speed and accuracy change? (Q2) ***How does the adaptation continue?*** In response to repeated exposure to the same (*a priori*) unfamiliar talker or accent, do the effects of exposure accumulate over time? Or alternatively, does adaptation begin anew each time?

Our current experiments built upon Xie et al. (2017) [15], who used a typical exposure-test design to examine L1-English listeners' adaptation to a Mandarin-accented word-final /d/-/t/ contrast in English. During exposure, Xie et al. (2017) used a lexical decision task where a target group of native listeners heard words produced by a Mandarin-accented talker. Importantly, some words included /d/ in a final position (e.g., *lemonade*), which is initially confusable with /t/ (i.e., *lemonade* sounding like *lemonate)*. In contrast, control group heard speech from the same Mandarin-accented talker but no /d/-final words. At test, both groups provided phoneme categorization responses to /d/-/t/ minimal pairs (e.g., *feed*-*feet*) produced by the same Mandarin-accented talker. Changes in the categorization responses between the two groups—elicited by the critical exposure to /d/-final words in the L2 accent (target group) or lack thereof (control group)—were taken as evidence for exposure-elicited adaptation to the L2 accented talker.

Extending this, we developed a new repeated testing paradigm in which listeners undergo three test blocks within a session (Fig. 1, top). This allows for finer-gained observations of adaptative changes in the recognition of unfamiliar speech inputs (Q1, Exp(eriment) 1). To examine long-term development of adaptive speech perception (Q2, Exp. 2), this repeated exposure-test design was administered five times spanning four weeks (Fig. 1, bottom). The test items were repeated across multiple blocks/days to increase the power of tracking trajectories of adaptation both within and across sessions.

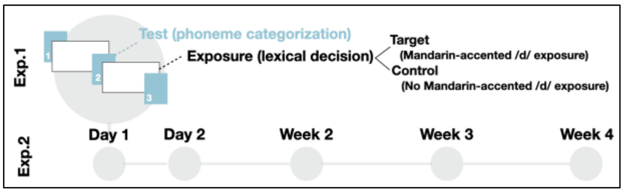

Exp.1
1
2
3
Test (phoneme categorization)
Exposure (lexical decision)
Target
(Mandarin-accented /d/ exposure)
Control
(No Mandarin-accented /d/ exposure)
Exp.2
Day 1
Day 2
Week 2
Week 3
Week 4

Figure 1 Exp. 1 consisted of three test blocks (10 items each) interspersed with two exposure blocks (90 items each) within a single session. Exp. 2 repeated the procedure from Exp. 1 five times: Day 1, one day later (Day 2), one week later (Week 2), two weeks later (Week 3), and three weeks later (Week 4).

2. EXPERIMENTS

Using the new repeated testing paradigm described above, we examined native English listeners' adaptation to L2- (Mandarin-)accented English.

2.1. Experiment 1 (One day)

Following the design of Xie et al. (2017), we compared recognition of L2-accented speech during test from listeners who were exposed to the critical /d/-final words vs. those who did not (the "target" vs. the "control" groups, respectively). Everything else was identical across the exposure conditions.

2.1.1. Participants

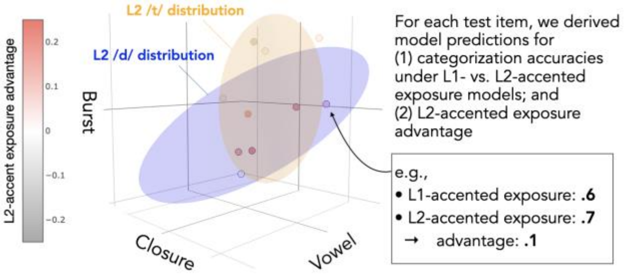
55 undergraduate students, all native speakers of American English, were recruited from the UC Irvine community and completed the experiment online via an online testing platform FindingFive. They were randomly assigned to the control exposure condition (n = 30) and target exposure condition (n = 25).

2.1.2. Stimuli

Both exposure and test stimuli were produced by a male native-Mandarin speaker of medium intelligibility in English. (The intelligibility was assessed by a separate norming study.) **Exposure stimuli** for the target group consisted of 90 English words (30 critical and 60 filler items) and 90 phonologically-legal nonwords. The critical items were all multi-syllabic words ending in /d/ (e.g., *lemonade*, *overload*). The exposure list for the control group was identical except that the 30 /d/-final critical items were replaced by 30 filler items. Filler words and nonwords did not contain any /d/ or /t/ sounds; no stop sounds other than /d/ appeared in word-final position. The exposure items were evenly distributed across the two exposure blocks.

Test stimuli consisted of five /d/- or /t/-final minimal pairs (e.g., *feed-feet*). The smaller number of test stimuli (8% of those used in Xie et al., 2017) was motivated by the current repeated testing paradigm. Because listeners go through multiple test blocks, presenting a large number of test items would run the risk of providing additional information about the target contrast and thereby neutralizing between-group differences (i.e., listeners in the control group may be able to adapt their recognition of the target contrast through the exposure to the test items alone). In addition, there is a growing awareness that the statistical information of the test tokens can attenuate (or potentially override) the effects of the exposure [16], [17]. Ideally, therefore, both exposure and test tokens should be sampled from the same natural cue distributions that characterize the target contrast.

Given these constraints, we applied a new method for sampling test tokens according to their acoustic-phonetic features. Specifically, we chose five minimal pairs based on: (a) Bayesian ideal observer simulations [18] and (b) human listener responses [15]. In (a), we simulated distributional learning of the underlying L2-accented categories along three cue dimensions critical for word-final /d/-/t/ contrast (Fig. 2). By contrasting the expected recognition accuracy after the L2- accented exposure (simulating the target group) vs. L1-accented exposure (simulating control participants who do not have exposure to L2-accented /d/), we computed the relative advantage expected from the L2-accented exposure (shown by the colorscale in Fig. 2). We then identified test items that were predicted to receive distinct categorization responses if native-English listeners indeed update their internal representations of /d/ vs. /t/ categories to match those in the L2-accented exposure input. In (b), we further narrowed our choices down to those test items that are *a priori* ambiguous to native listeners to avoid ceiling effects.



For each test item, we derived model predictions for
(1) categorization accuracies under L1- vs. L2-accented exposure models; and
(2) L2-accented exposure advantage

e.g.,
- L1-accented exposure: .6
- L2-accented exposure: .7
→ advantage: .1

Figure 2: Model predictions derived for the test items. The color of the dots indicates the degree of advantage expected from L2-(vs. L1-) accent exposure for individual test items.

2.1.3. Procedure

During exposure, participants completed a lexical decision (i.e., word vs. non-word) task. During test, participants provided 2AFC phoneme categorization responses to the minimal pair words (e.g., *feed-feet*). All participants completed an exit questionnaire about their language background and familiarity with L2 accents.

2.1.4 Results

Linear mixed effect logistic regression models were fit to the test data using the *lme4* package in R [19]. (Exposure) condition, test block, and category (/d/ vs. /t/) were examined as fixed effects. All models were specified with the maximal random effects structure justified by the design. Condition and category were sum contrast coded; test block was Helmert coded to compare (i) test2 vs. test1 and (ii) test3 to the mean of test1 and test2.

As predicted, /d/ words were recognized less accurately than /t/ words overall (β = -1.29, SE = .35, *z* = -3.66, *p* = .0002). A significant interaction between test block and category suggested that the improvement in test3 over the previous two tests was larger for /d/ and /t/ (β = .21, SE = .09, *z* = 2.26, *p* = .02). There was a marginally significant three-way interaction between condition, test block, and category (β = .18, SE = .09, *z* = 1.89, *p* = .06): Simple effects analyses revealed that only the target group significantly improved their recognition of /d/ (β = .52, SE =.15, *z* = 3.36, *p* = .0008).

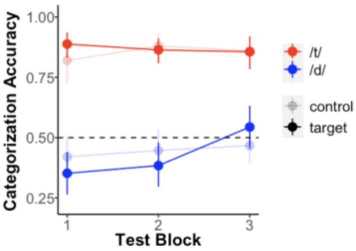


Figure 3: Results from Exp. 1 (1 day). Y-axis shows proportion of correctly categorized tokens in each exposure condition, over the three test blocks. Error bars represent 95% confidence intervals.

2.2. Experiment 2 (Five days spanning a month)

Exp. 2 repeated the same exposure-test session in Exp. 1 five times with one important change. In this experiment, the control condition used the stimuli identical to the target condition, except that they were produced by a native-English speaker (i.e., L1-accented exposure). This follows the design adopted by many influential studies on accent adaptation (e.g., [4]-[6], [20]), and it allowed us to keep constant the identities of words and the number of /d/ words heard across the two exposure conditions. Notably, neither of the target groups in Exp. 1 and 2 received exposure to Mandarin-accented /d/ (or /t/) sounds and therefore lacked the critical exposure needed for adaptation.

2.2.1. Participants

70 native speakers of American English, aged 18-45, were recruited via Prolific and completed the experiment online via FindingFive. Participants were randomly assigned to the L1-accented exposure condition (control, n = 38) and L2-accented exposure condition (target, n = 32).

2.2.2. Stimuli and procedure

The stimuli materials in both conditions were identical to those used in the target condition in Exp.1. Five exposure-test sessions, each identical to that of Exp. 1, were administered across five time points over the course of a month (Fig. 2).

2.3.3. Results

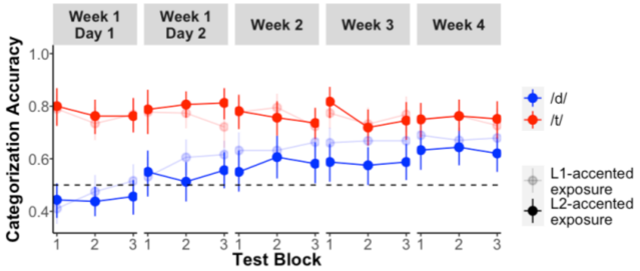


Figure 4: Results from Exp. 2 (multi-session spanning four weeks), collapsed across exposure conditions. Y-axis shows proportion of correctly categorized tokens within and across experimental sessions. Error bars represent 95% confidence intervals.

Analysis procedures were identical to those in Exp. 1, except that test block was recoded as a continuous variable (i.e., a total of 15 test blocks over five test sessions). This enabled us to examine whether test performance continued to change cumulatively across all test blocks. Both groups' overall performance improved significantly across time (β = .04, SE = .006, *z* = 6.63, *p* < .0001). Similar to Exp. 1, a test block-by-category interaction was found (β = .06, SE = .006, *z* = 9.54, *p* < .0001). This interaction was driven by an increase in recognition accuracy for /d/ (β = .10, SE = .008, *z* = 11.92, *p* < .0001) and a smaller decrease for /t/ (β = -.02, SE = .009, *z* = -1.95, *p* = .052). Unlike in Exp. 1, however, no three-way interaction was found: Both the target and the control groups in Exp. 2 improved their recognition of the L2-accented categories to a similar degree.

3. DISCUSSION

The current study combined two lines of work on native listeners' processing of L2-accented speech: (1) rapid adaptation in the initial moments of encountering an (*a priori*) unfamiliar talker/accent, and (2) longer-term accommodation of an accent through repeated exposure. Replicating previous findings, Exp. 1 demonstrated that listeners did indeed improve their perception within the first few minutes of exposure. The new repeated testing paradigm revealed detectable exposure effects *after* the second exposure block. This supports the idea that "rapid" adaptation still requires some amount of cumulative input from which listeners can extrapolate distributional structures of acoustic-phonetic cues characterizing the unfamiliar accent.

The results of Exp. 2 presented an intriguing puzzle. As expected, repeated exposure to a Mandarin-accented talker helped native listeners improve their recognition of the /d/ category of that talker without a corresponding decrease in the accuracy for the /t/ category, resulting in an increase in overall accuracy (from .61 to .69). We demonstrate, for the first time, that adaptation to naturally produced L2-accented categories continues to evolve over a month (see also [22], [23] for shorter-term adaptation to simpler stimuli).

Unexpectedly, control participants who received L1-accented exposure also achieved a comparable level of adaptation. By-item analysis confirmed that both groups responded to each of the five minimal pairs in a similar manner throughout the five sessions. In short, while the target group consistently benefited from exposure to L2-accented /d/ sounds over both timescales, the behavioural patterns of the control groups differed between Exp. 1 and Exp. 2.

What caused this difference? One might attribute it to differences in participant profiles, namely the campus population in Exp. 1 and the Prolific participants in Exp. 2. However, the fact that the participants had comparable recognition accuracy for /d/ and /t/ in the very first test block makes this an unlikely reason for the observed difference. Rather, it could be due to the number of /d/ tokens, which differed between the experiments. Recall that in Exp. 2, the control group received as many /d/ tokens as the target group did albeit from an L1-accented talker. This exposure, combined with the small number of minimal pair test items repeated across test blocks, may have drawn their attention to the /d/-/t/ contrast and led the control group to derive some strategic response patterns to improve their performance (e.g., I have responded /t/ to a similar item already, so I will respond /d/ to the current item).

Another, non–mutually exclusive, possibility is that *the test items alone* may have been sufficient to support the adaptation. i.e., Listeners in both the control and target conditions may have adapted to the accent-specific features through the exposure to the /d/-/t/ minimal pairs without labeling information. If this is indeed the case, it means that the L2-accented /d/ words heard during exposure did not provide any additional benefit to the target group. To address this, our future follow-up tests will include a condition in which listeners are tested on the same schedule as in the current study (= three test blocks / session * five sessions over a month) without any L2-accented exposure in between. If L1-listeners improve their recognition of the /d/-/t/ contrast in this condition, that would suggest that unsupervised adaptive changes in perception can happen faster and be more efficiently than has previously been believed.

More generally, the current data highlight the complexity of identifying the mechanism(s) for adaptive changes of recognition. Is there a single mechanism that explains adaptation across different timescales? Did participants in the control vs. target conditions use the same or different mechanism(s) to improve their recognition? By simply examining behavioral data aggregated over trials and subjects, as we did for the current study, we cannot know **how**—by what mechanism—the observed adaptive changes occurred.

To address this question, we have recently proposed a combinational framework (called ASP for "Adaptive Speech Perception") [24]. ASP instantiates how listeners' responses may adapt to exposure based on: (a) changes in auditory perception (e.g., low-level signal transformation and normalization), (b) changes in linguistic representations (e.g., distributional learning of phoneme categories), and (c) changes in decision making (e.g., updating of decision biases). Pursued in independent lines of work, these three mechanisms have so far rarely been contrasted or tested against each other using the same data.

ASP allows researchers to model adaptive changes in responses under any (combination) of the three mechanisms. This, in turn, allows us to select exposure and test items for human perception experiments in a targeted manner to more effectively contrast the predictions of the different mechanistic models. The new incremental testing paradigm developed in the current study will facilitate this approach: The improved ability to sample human responses to different types and amounts of exposure will increase the statistical power for reliable model comparisons. The paradigm, combined with model-based hypothesis-testing, thus holds promise to provide new insights into adaptive speech perception.

5. REFERENCES

[1] P. Adank and E. Janse, ‘Comprehension of a novel accent by young and older listeners.’, *Psychol Aging*, vol. 25, no. 3, pp. 736–740, Sep. 2010, doi: 10.1037/a0020054.

[2] V. Porretta, B. V. Tucker, and J. Järvikivi, ‘The influence of gradient foreign accentedness and listener experience on word recognition’, *J Phon*, vol. 58, pp. 1–21, Sep. 2016, doi: 10.1016/j.wocn.2016.05.006.

[3] M. M. Baese-Berk, D. J. McLaughlin, and K. B. McGowan, ‘Perception of non-native speech’, *Lang Linguist Compass*, vol. 14, no. 7, pp. 1–20, 2020, doi: 10.1111/lnc3.12375.

[4] X. Xie *et al.*, ‘Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker’, *J Acoust Soc Am*, vol. 143, no. 4, pp. 2013–2031, Apr. 2018, doi: 10.1121/1.5027410.

[5] A. R. Bradlow and T. Bent, ‘Perceptual adaptation to non-native speech.’, *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[6] C. Y. Tzeng, J. E. D. Alexander, S. K. Sidaras, and L. C. Nygaard, ‘The role of training structure in perceptual learning of accented speech’, *J Exp Psychol Hum Percept Perform*, 2016, [Online]. doi.org/10.1037/xhp0000260%5Cnhttp://

[7] C. M. Clarke and M. F. Garrett, ‘Rapid adaptation to foreign-accented English.’, *J Acoust Soc Am*, vol. 116, no. 6, pp. 3647–3658, 2004, doi: 10.1121/1.1815131.

[8] K. B. McGowan, ‘Social expectation Improves speech perception in noise’, *Lang Speech*, vol. 58, no. 4, pp. 502–521, Feb. 2015, doi: 10.1177/0023830914565191.

[9] A. Hanulíková, ‘Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups’, *PLoS One*, vol. 16, no. 10 October, Oct. 2021, doi: 10.1371/journal.pone.0259230.

[10] M. J. Witteman, A. Weber, and J. M. McQueen, ‘Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation’, *Atten Percept Psychophys*, vol. 75, no. 3, pp. 537–556, 2013, doi: 10.3758/s13414-012-0404-y.

[11] A. Weber, A. M. di Betta, and J. M. McQueen, ‘Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners’, *J Phon*, vol. 46, no. 1, pp. 34–51, 2014, doi: 10.1016/j.wocn.2014.05.002.

[12] B., Evans, and Iverson, P. ‘Plasticity in vowel perception and production: A study of accent change in young adults.’ *J. Accoustic Soc. Am. 121* (6), pp. 3814-3826, 2007, https://doi.org/10.1121/1.2722209

[13] Y. Zheng and A. G. Samuel, ‘The relationship between phonemic category boundary changes and perceptual adjustments to natural accents’, *J Exp Psychol Learn Mem Cogn*, vol. 46, no. 7, pp. 1270–1292, Jul. 2020, doi: 10.1037/xlm0000788.

[14] T. Bent and M. Baese-Berk, ‘Perceptual learning of accented speech’, *The Handbook of Speech Perception*, pp. 428–464, 2021, doi: 10.1002/9781119184096.ch16.

[15] X. Xie and E. B. Myers, ‘Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers’, *J Mem Lang*, vol. 97, pp. 30–46, 2017, doi: https://doi.org/10.1016/j.jml.2017.07.005.

[16] L. Liu, and T. F. Jaeger, ‘Inferring causes during speech perception’, *Cognition*, vol. 174, pp. 55–70. 2018. doi: 10.1016/j.cognition.2018.01.003

[17] Y. Zheng, and A. G. Samuel, ‘Flexibility and stability of speech sounds: The time course of lexically-driven recalibration’, *J Phon*, vol. 97, 2023. doi.org/10.1016/j.wocn.2023.101222

[18] M. Tan, X. Xie, and T. F. Jaeger, ‘Using rational models to interpret the results of experiments on accent adaptation’, *Front Psychol*, vol. 12, Nov. 2021, doi: 10.3389/fpsyg.2021.676271.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, ‘Fitting linear mixed-effects models using lme4’, *J Stat Softw*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[20] S. K. Sidaras, J. E. D. Alexander, and L. C. Nygaard, ‘Perceptual learning of systematic variation in Spanish-accented speech.’ *J Acoust Soc Am* **125**, 3306, 2009.

[21] X. Xie, L. Liu, and T. F. Jaeger, ‘Cross-talker generalization in the perception of non-native speech: A large-scale replication.’ *Journal of Experimental Psychology: General*, 150, e22–e56, 2021. https://doi.org/10.1037/xge0001039

[22] D. Saltzman and E. Myers, ‘Listeners are initially flexible in updating phonetic beliefs over time’, *Psychon Bull Rev*, 2021, doi: 10.3758/s13423-021-01885-1.

[23] M. J. Witteman, N. P. Bardhan, A. Weber, and J. M. McQueen, ‘Automaticity and stability of adaptation to a foreign-accented speaker’, *Lang Speech*, vol. 58, no. 2, pp. 168–189, Jun. 2015, doi: 10.1177/0023830914528102.

[24] X. Xie, T. F. Jaeger, and C. Kurumada, ‘What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review’. *Cortex*. To appear.