MECHANISMS OF GENERALIZATION FOR PHONETIC LEARNING OF ACCENTED SPEECH

Yevgeniy Vasilyevich Melguy and Keith Johnson

University of California, Berkeley
ymelguy@berkeley.edu

ABSTRACT

Listeners can rapidly adapt to an unfamiliar accent. Previous literature shows that such phonetic learning is often speaker-specific for speech sounds such as fricatives — it does not transfer to novel speakers, possibly because fricatives contain spectral properties that cue speaker identity. However, recent research shows that transfer to a novel sound contrast can occur within a single speaker, suggesting that there is room for analogous transfer of learning to a novel speaker with a sufficiently similar pronunciation. This study examines the generalization of phonetic learning for an atypical fricative pronunciation to novel speakers, focusing on the mechanisms that underlie such perceptual adaptation. Results from a set of experiments show that transfer of phonetic learning can occur to both novel speakers and novel phonetic contrasts, but that listeners may rely upon a more conservative mechanism when generalizing learning to novel speakers.

Keywords: speech perception, perceptual learning, phonetic learning, lexically-guided recalibration, accented speech

1. INTRODUCTION

Listeners show a remarkable ability to adapt to a novel accent, with a minimal amount of exposure generally sufficient to improve processing and/or comprehension [1, 2]. One example of such phonetic learning is lexically-guided recalibration of listeners' perceptual categories [3]. For instance, following exposure to a phonetically ambiguous pronunciation (/s/ = [s/f]) embedded in a word (e.g., mo?), listeners recalibrate their /s/ category such that they perceive more tokens along an /s/-/f/ phonetic continuum as /s/. Previous literature has suggested that this type of perceptual learning is often speaker-specific — it does not tend to generalize [4, 5]. However, recent literature suggests that there is some listener tolerance for phonetic mismatch between exposure and test contexts — listeners can generalize learning to a novel phonetic contrast containing the trained sound category [6]. Such results suggest that lexically-guided recalibration of phonetic categories is sufficiently robust to phonetic variation to facilitate speaker-independent adaptation to a given accent, as has been found in related literature on accent accommodation [2]. Crucially, listeners must be able to abstract over between-speaker differences in order to successfully learn the accent. Simultaneously, these must be able to pick up on regularities rising from interactions of the L1 and L2 sound systems. For instance, a common error by non-native speakers (NNS) of English involves the (inter-)dental frictives /θ/ and /ð/, because these sounds are typologically rare. NNS often substitute another sound that does exist in their L1 phoneme inventory (e.g., /f/, /s/, or /t/). However, even NNS with the same L1 background may realize a given sound in different ways [7], leading to perceptually and spectrally distinct realizations across speakers of the "same" accent. A study by [8] also found significant acoustic and articulatory variation in the production of these sounds across native American English speakers. So, achieving speaker-independent learning may require striking the right balance between perceptual flexibility and sensitivity: accommodating between-speaker variability while picking up on systematic phonetic patterns within a given accent.

To assess the nature of the mechanism underlying recalibration of phonetic categories, [6] tested listeners on a series of sound contrasts containing the trained phoneme. Results showed that learning was not contrast-specific, but generalized to a novel contrast that was perceptually similar to the training accent. This result is consistent with a phonetic learning strategy that involves expansion of a perceptual category into neighboring phonetic space, and suggests that the perceptual system utilizes a relatively coarse-grained learning mechanism in such situations. It is plausible that maintaining this kind of general learning strategy could help listeners achieve speaker-independent adaptation to a non-natively accented speech, which speech tends to be especially variable [9].

Existing literature does not provide a clear answer to the question of whether listeners rely on the same mechanisms for speaker-specific vs. speaker-independent phonetic learning [5, 10, 11]. Previous work has suggested that both relatively general and more specific mechanisms may be available to learners [12, 13]. The current study builds on recent work investigating the nature of the mechanism involved in perceptual recalibration or returning of phonetic categories [6]. We do so by testing cross-speaker generalization for multiple phonetic contrasts, in an attempt to understand what mechanisms underlie generalization of such learning.

2. EXPERIMENT 1

Experiment 1 followed the methods used in [6] while introducing a novel female and male speaker to test transfer of learning. Pilot testing found that intermixing tokens for all 3 speakers within a single test task prevented the detection of any training effect, contrary to results of [11, 14]. Therefore, the following experiments directly replicate the procedures in [6], where each group of participants only heard a single speaker.

2.1. Participants

The experiments in this study were conducted using a custom web-based program [15]. Initially, 255 participants were recruited via the Prolific online platform. All participants lived in the U.S. and reported being native speakers of American English with normal speech and hearing. Exclusion of participants who failed to meet experimental criteria (following [6]) resulted in the removal of 45 participants, leaving a total of 210 whose data were retained for analysis.

2.2. Stimuli

The same materials and procedures used in [6] were also used in this study. However, this experiment also included test continua produced by two novel speakers, recorded and processed using the same procedures, testing listeners on either (1) the male speaker from [6] who produced the training materials (age = 27), (2) a novel male speaker (age = 30), or (3) a novel female speaker (age = 25). All speakers were native speakers of American English, had grown up in California, and were living in California at the time of recording.

2.3. Procedure

Participants first had to complete a headphone check [16] . Participants in the accent exposure group then completed the same lexical decision task in [6] to familiarize them with the ambiguous accent (a pronunciation where /θ/ was replaced with [θ/s] in 20 critical words). All participants completed a categorization task using the same 4 /θ/-/s/ minimal-pair continua presented in [6]. Participants were randomly assigned to one of the three speakers. There were a total of 112 tokens (4 continua x 7 tokens x 4 repetitions per token). Tokens within each continuum were randomized and order of continua was counterbalanced across participants.

2.4. Analyses

Prior to analysis, trials with response times < 200 ms or > 2500 ms were removed, following [11, 14, 6]. Across the 3 speakers (familiar male, novel male, novel female), this resulted in removal of 736 trials (2.47% , 3.97%, and 2.72% of data per speaker). Data were analyzed via generalized linear mixed-effects regression modeling, using the lme4 package [17] in R [18]. Maximal random effect structure was used for each model where this did not result in convergence issues, with random slopes fitted for all within-participant main effects [19]. The significance of fixed effects and interactions between them was assessed using a Wald chi-squared test. An identical model was fitted for each group of listeners. Categorical variables were treatment-coded and included condition (training vs control, ref = control), /θ/ word position (word-initial vs. word-final, ref = word-final), and experiment half (trial block 1 vs. block 2, ref = block 1). Block was not included as a predictor in[6], but its inclusion is justified by recent studies showing that the recalibration effect can diminish or disappear over the course of testing [20]. A single numeric variable (continuum step, range = 1-7) was included as a centered numeric variable.

2.5. Results

Results for the familiar male speaker showed significant main effects of continuum step ($\chi^2(1) = 165.54$, $p < 0.001$), experimental group ($\chi^2(1) = 6.43$, $p < 0.05$), and block ($\chi^2(1) = 6.32$, $p < 0.05$). There were also significant 2-way interactions of continuum step by group ($\chi^2(1) = 5.53$, $p < 0.05$) and continuum step by block ($\chi^2(1) = 4.21$, $p < 0.05$). Finally, there were also significant 3-way interactions of step, group, and block ($\chi^2(1) = 6.73$,

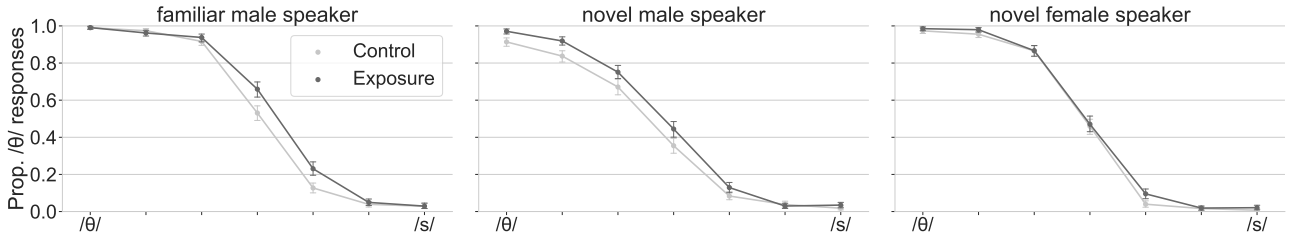

Figure 1: Categorization results for /θ/-/s/ minimal-pair phonetic continua. The effect of accent exposure was significant effect for the familiar male speaker and marginally significant effect for the novel male speaker.

$p < 0.01$), and step, word position, and block ($\chi^2(1) = 7.31$, $p < 0.01$). Crucially, the exposure group showed showed a significantly higher proportion (3.8%) of /θ/ responses (b = 0.86, SE = 0.34, z = 2.54, $p < 0.05$) compared to controls (see Fig. 1).

Results for the novel male speaker showed main effects of continuum step ($\chi^2(1) = 173.19$, $p < 0.001$), word position ($\chi^2(1) = 18.80$, $p < 0.001$), and trial block ($\chi^2(1) = 8.77$, $p < 0.01$), with a marginal effect of experimental group ($\chi^2(1) = 3.28$, $p = 0.070$). There were also significant 2-way interactions of continuum step by word position ($\chi^2(1) = 7.41$, $p < 0.01$) as well as word position by block ($\chi^2(1) = 8.16$, $p < 0.01$). There was also a significant 3-way interaction of step, word position, and block ($\chi^2(1) = 11.14$, $p < 0.001$). The exposure group showed a marginally significant (5.3%) increase in the proportion of /θ/ responses compared to controls (b = 0.57, SE= 0.31, z = 1.81, p = 0.070).

Results for the novel female speaker showed significant main effects of continuum step ($\chi^2(1) = 198.42$, $p < 0.001$) and of block ($\chi^2(1) = 18.91$, $p < 0.001$). There was also a significant 2-way interaction of group and block ($\chi^2(1) = 4.25$, $p < 0.05$). The exposure group showed showed a small numeric shift (1.8%) in their /θ/-/s/ category boundary (see Figure 1, but this difference was not significant (b = 0.13, SE = 0.26, z = 0.48, p = 0.63). There was also a significant decrease in the size of the training effect across experiment blocks (b = -0.92, SE = 0.44, z-score = -2.06, $p < 0.05$).

3. EXPERIMENT 2

Results of Experiment 1 replicated the key result in [6], and also provide tentative evidence of generalization of learning to a novel speaker of the same gender. This is consistent with previous phonetic recalibration studies showing that such learning for fricatives tends to resist cross-speaker generalization [5, 11], transfer only observed under highly similar exposure and test contexts.

The following set of experiments tests whether learning for the same ambiguous pronunciation (/θ/ = [θ/s]) can transfer to both a new speaker and a new phonetic contrast (/θ/-/ʃ/). Given that transfer of learning was observed within a single speaker in [6], it is possible that the same mechanism may facilitate cross-speaker generalization for this contrast. Given gender-based spectral differences in fricative realization [21, 22], we might expect a different pattern of results here, since the female speaker's /ʃ/ may be acoustically closer to the male speaker's /s/, allowing transfer for this contrast even though no transfer was observed in /θ/-/ʃ/ categorization.

The set of experiments presented below each utilize the same training materials and speakers as in Experiment 1. The test phase, however, involves categorization of /θ/-/ʃ/ minimal-pair continua, as in Experiment 2 of [6]. All procedures, materials, and analyses are otherwise identical to those of Experiment 1, a with a comparable number of participants (N=208 across the 3 speakers) and comparable data exclusion rates.

3.1. Results

For the familiar male speaker, model results revealed main effects of continuum step ($\chi^2(1) = 177.03$, $p < 0.001$) and block ($\chi^2(1) = 31.10$, $p < 0.001$). There was also a significant 2-way interaction of word position and block ($\chi^2(1) = 8.77$, $p < 0.01$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 6.00$, $p < 0.05$). Crucially, the effect of experimental group failed to reach significance ($\chi^2(1) = 2.38$, $p = 0.12$).

For the novel male speaker, model results revealed main effects of continuum step ($\chi^2(1) = 127.52$, $p < 0.001$), experimental group ($\chi^2(1) = 6.78$, $p < 0.01$), word position ($\chi^2(1) = 63.24$, $p < 0.001$), and trial block ($\chi^2(1) = 22.06$, $p < 0.001$). There were also significant 2-way interactions of step and word position ($\chi^2(1) = 5.73$, $p < 0.01$), of group and block ($\chi^2(1) = 10.43$, $p < 0.001$), and of word position and

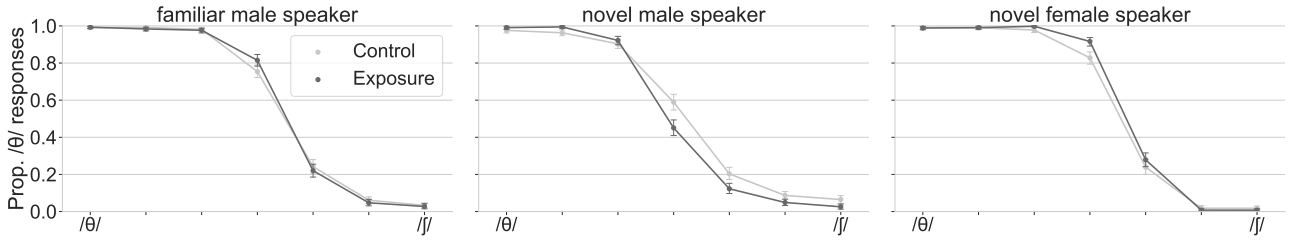

Figure 2: Categorization results for /θ/-/ʃ/ minimal-pair phonetic continua. A significant effect of accent exposure was observed for the novel male and female speakers, but not for the familiar male speaker.

block ($\chi^2(1) = 15.66$, $p < 0.001$). There were also 3-way interactions of step, group, and word position ($\chi^2(1) = 9.77$, $p < 0.01$), of step, word position, and block ($\chi^2(1) = 6.28$, $p < 0.05$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 5.36$, $p < 0.05$). The training effect showed up as a significant decrease (-3.2%) in the proportion of /θ/ responses in the data (b = -0.90, SE =0.35, z = -2.60 $p < 0.01$), but this effect shrank in size over the course of the task: the -5.1% shift seen in the first half of trials decreased to -1.3% in the second half (b = 0.85, SE = 0.26, z = 3.23, $p < 0.01$).

For the novel female speaker, results showed main effects of continuum step ($\chi^2(1) = 112.44$, $p < 0.001$), experimental group ($\chi^2(1) = 5.01$, $p < 0.05$), and trial block ($\chi^2(1) = 5.21$, $p < 0.05$), with a significant interaction of step and word position ($\chi^2(1) = 5.66$, $p < 0.05$). Training resulted in a significant increase (1.8%) in the proportion of /θ/ responses (b = 1.00, SE = 0.45, z = 2.24, $p < 0.05$)

4. GENERAL DISCUSSION

Overall, these results provide evidence that recalibration of phonetic categories can affect perception of novel speakers, consistent with previous work [5, 23, 11]. This study builds on previous work by demonstrating that it is possible for recalibration of phonetic categories to transfer to both novel speakers *and* novel phonetic contrasts involving the trained target sound. The pattern of results suggests that listeners may rely on a more conservative perceptual strategy when generalizing learning to novel speakers, compared to the mechanisms of adaptation for a single speaker reported in earlier work [6].

Experiment 1 replicated previous work [6], finding perceptual learning for an atypical /θ/=[θ/s] pronunciation for the familiar male speaker, and also showing a trend for generalization to a novel male—but not a novel female—speaker. Experiment 2 failed to find perceptual learning for the familiar speaker, but found that training did affect perception of both a novel male and female speaker. Results here show an opposite pattern across speakers, where listeners tested on the female speaker showed a great likelihood to classify ambiguous [θ/s] tokens as exemplars of /θ/, while those tested on the male speaker were *less* tolerant of such atypical pronunciations.

Overall, this pattern of results suggests that transfer of learning is constrained by a relatively targeted mechanism. Crucially, we do not find support for the category expansion mechanism reported in recent work [6], as lack of learning in the familiar speaker condition in Exp.2 fails to replicate their result. While results of Exp.1 are explicable in terms of between-speaker similarity, interactions between speaker gender, sound contrast, and transfer of learning in Exp.2 are puzzling and not clearly aligned with previous literature where cross-speaker and cross-gender learning has been observed [5]. The negative categorization shift for the novel male speaker in Exp.2, where listeners were less likely to perceive /θ/ following accent exposure, is particularly puzzling, as such a learning effect has not been reported in previous phonetic recalibration literature. A fuller discussion of these points, including acoustic analyses, may be found in Ch. 4 of [24], with all study materials publically available at [25].

In summary, these results suggest that generalization of phonetic learning is relatively constrained, consistent with previous work [5, 4, 10, 11]. Nonetheless, learning did transfer to both novel female and male speakers under particular testing conditions, challenging earlier work suggesting that phonetic learning for fricatives is entirely speaker-specific [4]. The mixed results in this study are not entirely surprising given the conflicting pattern of results in the broader literature, where generalization following single-speaker accent exposure has been found in some studies [26] but not others [2]. Future work may shed additional light on the the precise factors modulating such cross-speaker transfer of learning.

5. ACKNOWLEDGEMENTS

This research was supported in part by a Dissertation Completion Fellowship and a National Science Foundation (NSF) Graduate Research Fellowship to Y.V.M. (Grant No. 1752814). We thank Isaac Bleaman, Terry Regier, and Frederic Theunissen for helpful comments.

6. REFERENCES

[1] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[2] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[3] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cognitive Psychology*, vol. 47, no. 2, pp. 204–238, 2003.

[4] F. Eisner and J. M. McQueen, "The specificity of perceptual learning in speech processing," *Perception & Psychophysics*, vol. 67, no. 2, pp. 224–238, 2005.

[5] T. Kraljic and A. G. Samuel, "Perceptual learning for speech: Is there a return to normal?" *Cognitive Psychology*, vol. 51, no. 2, pp. 141–178, 2005.

[6] Y. V. Melguy and K. Johnson, "Perceptual adaptation to a novel accent: Phonetic category expansion or category shift?" *The Journal of the Acoustical Society of America*, vol. 152, no. 4, pp. 2090–2104, 2022.

[7] A. Seibert, "A sociophonetic analysis of L2 substitution sounds of American English interdental fricatives," Master's thesis, Southern Illinois University at Carbondale, 2011.

[8] Y. Melguy, "Strengthening, weakening and variability: The articulatory correlates of hypo-and hyper-articulation in the production of english dental fricatives," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[9] T. Wade, A. Jongman, and J. Sereno, "Effects of Acoustic Variability in the Perceptual Learning of Non-Native-Accented Speech Sounds," *Phonetica*, vol. 64, no. 2-3, pp. 122–144, 2007.

[10] T. Kraljic and A. G. Samuel, "Perceptual adjustments to multiple speakers," *Journal of Memory and Language*, vol. 56, no. 1, pp. 1–15, 2007.

[11] E. Reinisch and L. L. Holt, "Lexically guided phonetic retuning of foreign-accented speech and its generalization." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 2, pp. 539–555, 2014.

[12] R. Schmale, A. Cristia, and A. Seidl, "Toddlers recognize words in an unfamiliar accent after brief exposure: Brief exposure to an unfamiliar accent," *Developmental Science*, vol. 15, no. 6, pp. 732–738, 2012.

[13] D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel." *Psychological Review*, vol. 122, no. 2, pp. 148–203, 2015.

[14] E. Reinisch, A. Weber, and H. Mitterer, "Listeners retune phoneme categories across languages." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 1, pp. 75–86, 2013.

[15] K. Johnson, "Speech perception on the web," 2022. [Online]. Available: https://github.com/keithjohnson-berkeley/perception_on_the_web

[16] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," *arXiv:1406.5823 [stat]*, 2014.

[18] R. C. Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2022.

[19] D. J. Barr, "Random effects structure for testing interactions in linear mixed-effects models," *Frontiers in Psychology*, vol. 4, 2013.

[20] L. Liu and T. F. Jaeger, "Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 12, pp. 1562–1588, 2019.

[21] V. A. Mann and B. H. Repp, "Influence of vocalic context on perception of the [sh]-[s] distinction," *Perception & Psychophysics*, vol. 28, no. 3, pp. 213–228, 1980.

[22] E. A. Strand and K. Johnson, "Gradient and Visual Speaker Normalization in the Perception of Fricatives," in *Natural Language Processing and Speech Technology*, D. Gibbon, Ed. De Gruyter, 1996, pp. 14–26.

[23] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychonomic Bulletin & Review*, vol. 13, no. 2, pp. 262–268, 2006.

[24] Y. V. Melguy, "Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent," Ph.D. dissertation, University of California, Berkeley, 2022.

[25] Y. Melguy, "Mechanisms of perceptual learning," 2022. [Online]. Available: https://github.com/ymelguy/mechanisms_of_perceptual_learning

[26] X. Xie and E. B. Myers, "Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers," *Journal of Memory and Language*, vol. 97, pp. 30–46, 2017.