GENERALIZED PERCEPTUAL ADAPTATION TO SECOND-LANGUAGE (L2) SPEECH: VARIABILITY, SIMILARITY AND INTELLIGIBILITY

Ann R. Bradlow[1], Adrianna M. Bassard[2], and Ken A. Paller[2]

[1]Department of Linguistics, [2]Department of Psychology, Northwestern University, Evanston, IL, USA
abradlow@northwestern.edu, AdriannaBassard2022@u.northwestern.edu, kap@northwestern.edu

ABSTRACT

Recent work suggests that generalization of perceptual adaptation to L2 speech depends on similarity between training and test talkers as well as on high-variability training. We explored this proposal for both talker- and accent-independent adaptation with a round-robin design involving multi-talker testing following single-talker (ST) and multiple-talker (MT) training. Test sentences were produced by four L2 English talkers from different L1 backgrounds. Training involved four single-talker (ST) and four multiple-talker (MT) conditions with all talkers serving as both test and training talkers. On average, greater intelligibility improvement resulted from MT than ST training; however, some training-test talker combinations also showed generalized adaptation following ST training. Critically, variation in training talker baseline intelligibility related to variation in adaptation and generalization. Together with prior work, these data suggest that generalized adaptation to L2 speech is sensitive to a combination of high-variability training, training-test talker similarity, and baseline intelligibility of the training talker(s).

Keywords: perceptual adaptation, speech intelligibility, L2 speech.

1. INTRODUCTION

Both human and computer listeners often struggle to recognize speech by second-language (L2) talkers [1-4]. However, prior work has shown substantial exposure-induced improvement in L2 speech recognition for human listeners (for a review see, [5]) indicating remarkably flexible human speech perception in response to extensive variation in speech production. A key issue for this research is to identify the exposure conditions that lead to the most extensive generalization. Optimal adaptation would generalize from a small set of training stimuli to improved speech recognition accuracy for speech by novel (i.e., untrained) L2 talkers from both trained and novel accents. In addition to the benefits for enhancing speech communication across a language barrier, this research enterprise provides a window into the processes and representations that support perceptual constancy in the face of extensive speech variation.

One approach emphasizes exposure to variability during training as a means of directing listener attention toward systematic and away from random variation. Research in this approach has successfully demonstrated cross-talker [6-8] and cross-accent [9,10] generalization after exposure to multiple talkers of a single accent (e.g., Chinese-accented English) or multiple talkers from different accent groups (e.g., Chinese-accented, Spanish-accented, Korean-accented, and Hindi-accented English), respectively. In contrast, learning following exposure to a single talker or accent was limited to the trained talker or accent, respectively [e.g., 6].

Other work has emphasized similarity between training and test stimuli rather than exposure to variability [11-13]. For example, [13] demonstrated cross-talker generalization following both multiple- and single-talker training with a design that involved 20 unique training-test talker combinations with all talkers appearing in both training and test phases. This design effectively disentangled training format (single-talker versus multiple-talker) from the specific training and test talkers, and thereby revealed variation in generalization across different training-test talker pairs with some single-talker conditions leading to as much cross-talker/within-accent generalization as multiple-talker training. This finding is consistent with a similarity-based account according to which generalization of adaptation beyond training stimuli depends on sufficient overlap between training and test talkers in their patterns of phonetic realization of phonological contrasts [11-13]. In this view, any observed benefit of high variability over low variability training is not due to exposure to variability per se; instead, the benefit of training set variability is the increased likelihood of exposure to training stimuli that are sufficiently similar to the test stimuli to facilitate cross-talker generalization.

The present study explored the variability-based and similarity-based approaches for both cross-talker and cross-accent generalization with a design that like [13] and [10] but unlike [9] disentangled training format (single-talker, ST, versus multiple-talker, MT) from the specific talkers presented in the training and test phases. Moreover, like [9] and [13] but unlike

[10] the present study investigated both talker-general and accent-general adaptation with sentences rather than words as the training and test stimuli. This design feature is motivated by the fact that phrase-level and other supra-segmental and supra-lexical aspects of phonetic structure are salient, language-general features of naturally produced L2 speech [14, 15, 16] and therefore potentially important cues for perceptual adaptation particularly across accents.

2. METHOD

The overall design for this study involved a training phase followed by a test phase with an 11-12-hour delay between training and test. This delay was in anticipation of a subsequent study (not reported here) involving sleep consolidation. Eight different training conditions (between-participants) were followed by an identical multiple-talker, multiple-accent sentence-in-noise recognition test. A round-robin arrangement of training and test talkers allowed us to compare perceptual adaptation to L2 English across various training-test talker combinations (n=16) following both single-talker (ST) and multiple-talker (MT) training. Importantly, the four talkers involved in the round robin all came from different L1 backgrounds, allowing examination of both talker-specific and talker-general/accent-general adaptation to L2 speech.

2.1. Participants

A total of 195 first-language (L1) American English listeners participated in this study. All participants were between 18-35 years and self-reported as having no deficits in speech, language, or hearing, and as having normal or corrected-to-normal vision.

2.2. Materials and procedure

A total of 300 sentence recordings were downloaded from an open-access corpus of L2 speech that includes recordings from over 100 L2 talkers from over 20 L1 backgrounds [17]. For the present study, we compiled a set of simple sentences (e.g., “A towel is near the sink.”) from four L2 talkers from four different L1 backgrounds (75 sentences per talker), two males (L1 Brazilian Portuguese and L1 Spanish), and two females (L1 Farsi and L1 Turkish). These talkers were selected based on informal, subjective judgements by the authors as clearly L2-accented with moderate-to-good comprehensibility.

In both the training and test phases, participants listened over headphones or earbuds to sentence recordings that had been digitally mixed with speech-shaped noise at a fixed signal-to-noise ratio of 0 dB (i.e., the speech and noise were presented with equal loudness). The sentences were presented one at a time with no possibility of repetition. Participants typed what they heard using the computer keyboard before advancing to the next sentence. No feedback was provided in either the training or test phase.

At test, participants were presented with 15 sentences by each of the four talkers (total = 60 sentences). The talker-sentence pairings and trial order were held constant across all training conditions. Participants were randomly assigned to one of eight training conditions, four single-talker (ST) and four multiple-talker (MT) (n=20-22 per condition). All training conditions included the same 60 sentences (different from the test sentences). An additional untrained control group (n=27) took the multiple-talker test without any prior training phase.

In the ST training conditions, all 60 sentences were produced by one of the four talkers. Thus, at the test phase, listeners in ST training conditions encountered one trained talker and three novel talkers. In the MT training conditions, three of the four talkers produced 20 sentences each, while the fourth talker was excluded from the training set. Thus, at the test phase, listeners in MT training conditions encountered three trained talkers and one novel talker.

All sentence transcriptions were scored using an open-source automated scoring tool, Autoscore [18], which counts a sentence as correctly (score=1) or incorrectly (score=0) recognized if and only if the transcription exactly matches the talker’s script. Obvious spelling errors or homophones of intended words counted as correct. Sentence-level rather than word-level intelligibility was the dependent variable (DV) in all analyses because words in sentences are not equally independent of each other (due to sentence-level structure). In keeping with prior work, we chose intelligibility (word recognition accuracy) as the DV rather than comprehensibility or accentedness (see [19] for discussion on this point).

For analysis, proportional scores were log-odd transformed: $\log((p)/(1-p))$ where p is the intelligibility score ranging from 0 to 1. Separate ST and MT analyses were conducted because of the different balance of trained versus novel talkers encountered in the test following ST versus MT training (1 trained and 3 novel for ST versus 3 trained and 1 novel for MT). Within each analysis (ST or MT), t-tests with Bonferroni correction for multiple comparisons (n=16) compared performance at test for each training talker-test talker pairing relative to that test talker’s baseline intelligibility (control condition). Additionally, relative entropy (Kullback-Leibler divergence) of the log-odd transformed test score means was calculated for each training condition (n=8) with respect to the control condition. This information theoretic metric is a non-zero

number that quantifies "informativity" (or suprisal) of the distribution of scores across the four test talkers relative to the untrained control condition. Relative entropy of 0 would indicate no divergence from control to test (i.e., no adaptation); while, higher and lower relative entropy values indicate more or less divergence from baseline, respectively (i.e., more or less improvement in speech recognition accuracy relative to untrained controls). The equation for relative entropy is shown in (1), where P(x) is average intelligibility score for a given test talker, X is the number of test talkers within each training condition (4), and Q(x) is the given test talker's baseline intelligibility score.

(1) $$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

3. RESULTS

	Test talker			
	BRP	FAR	SPA	TUR
Control	0.784 (0.022)	0.722 (0.025)	0.712 (0.021)	0.545 (0.025)
Single-talker training (ST)				
BRP	**0.843 (0.023)**	0.725 (0.032)	0.727 (0.03)	0.572 (0.032)
FAR	0.886 (0.016)	**0.778 (0.023)**	0.767 (0.019)	0.599 (0.02)
SPA	0.852 (0.024)	0.723 (0.035)	**0.725 (0.035)**	0.581 (0.034)
TUR	0.794 (0.026)	0.676 (0.039)	0.708 (0.029)	**0.566 (0.031)**
Multiple-talker training (MT)				
noBRP	**0.854 (0.024)**	0.76 (0.025)	0.727 (0.029)	0.626 (0.025)
noFAR	0.827 (0.021)	**0.749 (0.035)**	0.714 (0.033)	0.609 (0.034)
noSPA	0.817 (0.028)	0.739 (0.033)	**0.727 (0.032)**	0.566 (0.031)
noTUR	0.878 (0.014)	0.794 (0.016)	0.778 (0.021)	**0.617 (0.024)**

Table 1: Average proportion of sentences correctly recognized for each test talker in untrained control, single-talker (ST), and multiple-talker (MT) training conditions. For ST, talker-specific ("old") trials are in bold. For MT, novel talker trials are in bold. Std. error in parentheses.

Table 1 shows average sentence recognition accuracy across all participants in the untrained control condition, the four ST conditions, and the four MT training conditions. The speech recognition accuracy scores are broken down by test talker (columns). For the ST training format, trials with matched training and test talkers are shown in bold. L1 codes for the ST training and test talkers are BRP = Brazilian Portuguese, FAR = Farsi, SPA = Spanish, TUR = Turkish. For MT training, condition codes indicate the excluded talker (all three of the other talkers were included): noBRP = Brazilian Portuguese excluded, noFAR = Farsi excluded, noSPA = Spanish excluded, noTUR = Turkish excluded. Test trials with the novel talker are shown in bold.

Figure 1 shows sentence recognition accuracy for all ST (top) and MT (bottom) training conditions by test talker. In each plot, the line indicates baseline intelligibility for the test talkers (untrained control condition). Boxes match bold entries in Table 1, i.e., indicate talker-specific ('old' talker) for ST and talker-generalization ('new' talker) for MT.

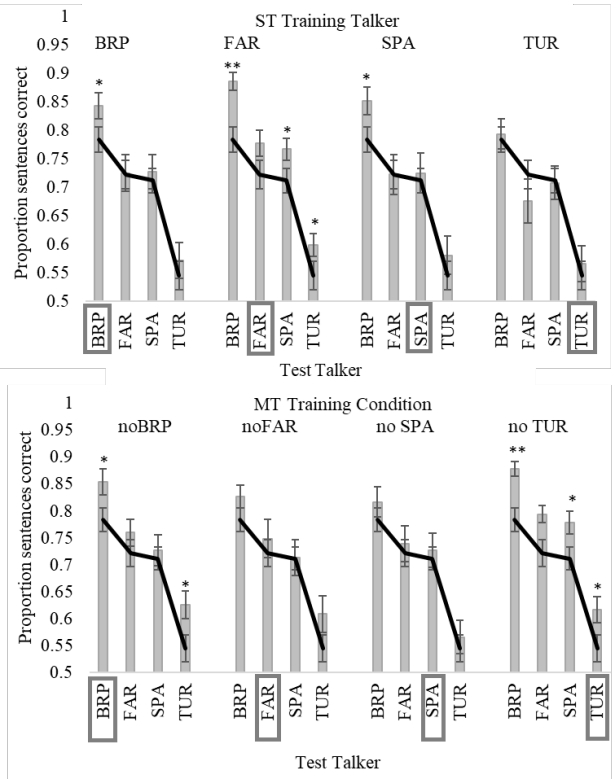

Figure 1: Sentence recognition accuracy for all ST (top) and MT (bottom) training conditions. Error bars are SE. Lines indicate baseline intelligibility for each test talker. Boxes indicate 'old' talker trials for ST and 'new' talker trials for MT. For pair-wise comparisons, * = p<.05, ** = p<.003 (Bonferroni correction for 16 comparisons).

Within the ST training format we see variation across the 16 combinations of training and test talkers. Significant training-to-test improvement (at the Bonferroni corrected level of p<.003) emerged for one test talker, BRP, following ST training with one talker, FAR. All other differences from baseline intelligibility failed to reach significance at this level although several others were significant at the less conservative level of p<.05 (FAR training for SPA and TUR test talkers, SPA training for BRP test talker, and BRP training for BRP test talker). Similarly, within the MT training format, we see

substantial variation across the 16 combinations of training condition and test talker. Pair-wise comparisons showed significant training-to-test improvement (at the Bonferroni corrected level of $p<.003$) for the BRP test talker following the MT training condition that excluded the TUR talker. All other differences from baseline intelligibility failed to reach significance at this level although several others were significant at the less conservative level of $p<.05$ (noTUR training for SPA and TUR test talkers; noBRP training for BRP test talker).

Figure 2 shows relative entropy for each training condition grouped by training format (ST vs. MT). Each data point represents overall divergence of test from baseline scores for the four test talkers in the indicated training condition. Lines connect ST and MT training conditions that include (ST) or exclude (MT) the talker indicated, e.g., the FAR-noFAR line connects the ST condition with FAR as the training talker and the MT condition that excluded this talker, noFAR. The overall pattern reveals slightly higher average relative entropy following MT than ST training. However, the individual training conditions vary in the direction of change from ST to MT. In particular, relative entropy decreases for the MT condition that excludes the "best" ST trainer (FAR to noFAR) and increases for the MT condition that excludes the "worst" ST trainer (TUR to noTUR).

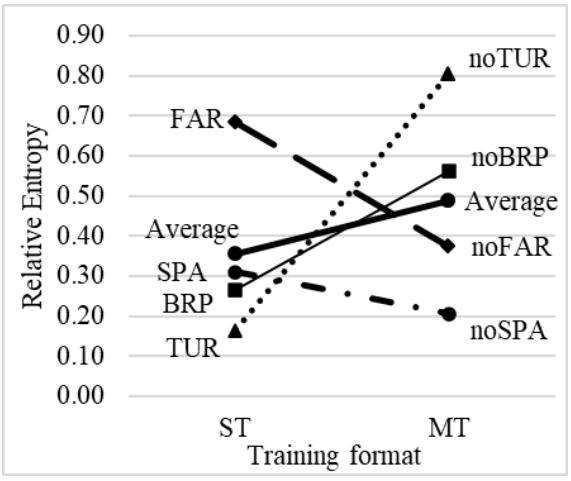

Figure 2: Relative entropy (Kullback-Leibler divergence) for each training condition grouped by training format (ST vs. MT). See text for more explanation.

4. DISCUSSION

The present study revealed exposure-induced cross-talker, cross-accent improvement in L2 English recognition by L1 English listeners following both ST and MT training. The extent of adaptation varied across training and test talker pairings; however, in both formats we found evidence of significant cross-accent as well as cross-talker generalization.

The relative entropy (Kullback-Leibler divergence) analysis showed that average relative entropy for MT training was higher than for ST training. This reflects greater divergence from baseline of test scores across the four test talkers following MT training than following ST training. Moreover, Figure 2 shows a cross-over effect for the least (TUR) and most (FAR) effective ST training talkers. Specifically, the upward sloping TUR-noTUR line indicates that exclusion of the least effective ST training talker (TUR) from the MT format (noTUR) benefitted adaptation, while the downward sloping FAR-noFAR line indicates that exclusion of the most effective ST training talker (FAR) detracted from adaptation.

The patterns emerging in this dataset suggest a possible role for baseline intelligibility in determining the extent of generalized perceptual adaptation to L2 speech. Listeners showed most consistent improvement across all training conditions for the test talker with the highest baseline intelligibility, BRP. Conversely, the talker with the lowest baseline intelligibility, TUR, was the least effective training talker. Variation in baseline intelligibility cannot account for all aspects of the present dataset, e.g., SPA and FAR have similar baseline intelligibility but they do not show qualitatively or quantitively equivalent performance as either training or test talkers. Nevertheless, association of the training and/or test talker's baseline intelligibility with extent of generalized adaptation to L2 speech noted above (i.e., for BRP and TUR) is consistent with other research on lexically-guided perceptual learning for speech which has demonstrated a crucial role for lexical knowledge in exposure-induced recalibration of phonetic category boundaries [20, 21 and many others]. If word recognition accuracy is too low (as for a low-intelligibility L2 talker) then lexical knowledge may not be available to guide the mapping of phonetic variation to linguistically meaningful categories, and consequently adaptation to L2 speech may be constrained for both talker-specific and talker/accent-general adaptation. In these cases, more exposure or other source of lexical access (e.g., through written medium) may be needed for effective adaptation. Conversely, if word recognition accuracy is high then lexically-guide recalibration of phonetic categories can presumably proceed readily.

In conclusion, together with prior work [5-13], the present study suggests that variability, similarity, and intelligibility of both training and test talkers are all relevant for determining the balance of specificity and generality of perceptual adaptation to L2 speech. Future research should directly test the separate and combined effect of each of these factors on generalized perceptual adaptation to L2 speech.

7. REFERENCES

[1] Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97.

[2] Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency on intelligibility of Chinese-accented English. *Language and Speech*, 47(2), 139–154.

[3] McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, & Psychophysics*, 80(6), 1559–1570.

[4] Harwell, D. (2018). The Accent Gap. *Washington Post*, 1–11.

[5] Bent, T. & Baese-Berk, M. (2021). Perceptual learning of accented speech. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni, eds. The Handbook of Speech Perception, 2nd ed. John Wiley & Sons, Inc., pp. 428–464.

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

[7] Tzeng, CY, Alexander, JD, Sidaras, SK, and Nygaard, LC (2016). The role of training structure in perceptual learning of accented speech. *J. Exp. Psychol. Hum. Percept. Perform*. 42, 1793–1805.10.1037/xhp0000260

[8] SK Sidaras, JED Alexander, LC Nygaard (2009) Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America* 125(5), 3306-3316.

[9] Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America,* 133(3), EL174-EL180.

[10] Alexander, J., and Nygaard, LC (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, 145(6), 3382.

[11] X Xie, EB Myers (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language* 97, 30-46

[12] X Xie, FS Earle, EB Myers (2018) Sleep facilitates generalization of accent adaptation to a new talker. *Language, Cognition and Neuroscience* 33 (2), 196-210, https://doi.org/10.1121/1.5110302

[13] Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, 150(11), e22–e56.

[14] Baker, R. E., Baese-Berk, M., Bonnasse-Gahot, L., Kim, M., Van Engen, K. J., & Bradlow, A. R. (2011). Word durations in non-native English. *Journal of phonetics*, 39(1), 1-17.

[15] Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, 138(3), EL223-EL228.

[16] Bradlow, A. R. (2022). Information encoding and transmission profiles of first-language (L1) and second-language (L2) speech. *Bilingualism: Language and Cognition*, 25(1), 148-162.

[17] Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar

[18] Borrie, S.A., Barrett, T.S., & Yoho, S.E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *Journal of Acoustical Society of America*, 145, 392-399.

[19] Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68(1), 115-146.

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

[21] Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, 13, 262–268.