Short-term implicit voice-learning leads to a Familiar Talker Advantage: The role of encoding specificity

Julie Case[a)]
Department of Communicative Sciences and Disorders, New York University, 665 Broadway, 9th floor, New York, New York 10012, USA
julie.case@nyu.edu

Scott Seyfarth
Department of Linguistics, Ohio State University, 1712 Neil Avenue, Oxley Hall, Columbus, Ohio 43210, USA
seyfarth.2@osu.edu

Susannah V. Levi
Department of Communicative Sciences and Disorders, New York University, 665 Broadway, 9th floor, New York, New York 10012, USA
svlevi@nyu.edu

Abstract: Whereas previous research has found that a Familiar Talker Advantage—better spoken language perception for familiar voices—occurs following explicit voice-learning, Case, Seyfarth, and Levi [(2018). J. Speech, Lang., Hear. Res. **61**(5), 1251–1260] failed to find this effect after *implicit* voice-learning. To test whether the advantage is limited to explicit voice-learning, a follow-up experiment evaluated implicit voice-learning under more similar encoding (training) and retrieval (test) conditions. Sentence recognition in noise improved significantly more for familiar than unfamiliar talkers, suggesting that short-term implicit voice-learning can lead to a Familiar Talker Advantage. This paper explores how similarity in encoding and retrieval conditions might affect the acquired processing advantage.

© 2018 Acoustical Society of America
[RS]
Date Received: August 18, 2018 **Date Accepted:** November 12, 2018

1. Introduction

The Familiar Talker Advantage is a phenomenon in which listeners are better at understanding speech produced by familiar voices. The advantage has been typically reported following explicit and extensive voice-identification training in which similar conditions (e.g., pre-recorded stimuli) are used for both the learning and the spoken language processing tasks (Levi, 2015; Levi *et al.*, 2011; Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994; Yonan and Sommers, 2000). A few studies have demonstrated the Familiar Talker Advantage without identification training. However, these studies have involved either explicit voice recognition during training with pre-recorded stimuli (Yonan and Sommers, 2000), or else extensive implicit exposure to the voice of a spouse or close friend (Johnsrude *et al.*, 2013; Souza *et al.*, 2013).

As voice-learning typically occurs implicitly in everyday life, Case *et al.* (2018) examined whether child and adult listeners would demonstrate a Familiar Talker Advantage following short-term implicit voice-learning. In that study, familiarization with a talker's voice occurred through face-to-face interactions with one of two talkers. The Familiar Talker Advantage was assessed by looking for improvement between baseline and post-learning in participants' ability to repeat pre-recorded sentences presented in noise. Unexpectedly, neither child nor adult participants demonstrated significantly more improvement on the sentence recognition task for the familiar talker.

While short-term implicit perceptual attunement—in which listeners' speech processing improves as a result of experience with a specific perceptual task—is well established in the speech perception literature (Bent *et al.*, 2009; Bradlow and Bent, 2008; Kreitewolf *et al.*, 2017; Van Engen, 2012), the Familiar Talker Advantage crucially involves a processing advantage that *transfers* from one task to another

[a)]Author to whom correspondence should be addressed.

(e.g., voice identification transfer to better spoken word recognition). Because we previously failed to find a Familiar Talker Advantage following implicit voice-learning with a transfer task, yet implicit attunement is well-attested when the task does not change, this raises the question of whether the Familiar Talker Advantage might be moderated by other kinds of task-to-task similarities, such as stimulus presentation mode.

1.1 The current study

The lack of a Familiar Talker Advantage could result from a variety of differences in the design used in Case *et al.* (2018) compared to the designs of previous studies (Levi, 2015; Levi *et al.*, 2011; Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994; Yonan and Sommers, 2000), including the amount of exposure to the voices (short-term versus long-term exposure), the type of exposure (explicit training versus implicit learning), or differences in how stimuli were presented during the exposure and testing phases. The current study explores this latter possibility. Research on encoding specificity suggests that individuals access newly learned information more easily when the learning conditions are similar to the testing conditions (Tulving and Thomson, 1973). It is therefore possible that differences in how voice information was learned (i.e., through in-person, live interactions) versus retrieved during the experimental task (i.e., with pre-recorded stimuli over headphones) may have impacted participant performance in our previous study. With the exception of Souza *et al.* (2013) and Johnsrude *et al.* (2013), all previous experiments on the Familiar Talker Advantage have used the same presentation mode (pre-recorded stimuli) for both voice-learning and spoken language processing tasks (Levi, 2015; Levi *et al.*, 2011; Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994; Yonan and Sommers, 2000). To evaluate whether short-term implicit voice-learning can also generate a Familiar Talker Advantage in this context, we conducted a follow-up study to Case *et al.* (2018) in which stimulus presentation was more similar during encoding (implicit voice-learning) and retrieval (sentence recognition). In addition to providing a new test of the Familiar Talker Advantage after short-term implicit voice-learning, this experiment can be combined with our previous results to explore the relevance of context similarity.

2. Methods

2.1 Participants

Thirty-one adults, ages 18–33 yr (mean age: 22.6 yr, 10 male, 21 female), were included in the final data analysis. Participants were native speakers of American English with no reported history of speech-language or hearing impairments. All participants passed a hearing screening at 25 dB sound pressure level at 500, 1000, 2000, and 4000 Hz using a portable Earscan3 Screening Audiometer (Micro Audiometrics, Murphy, NC). Seven additional participants were eliminated from the study due to living outside the U.S. before the age of 1 ($n = 1$), being a non-native speaker of American English ($n = 1$), not attending the second day of the study ($n = 2$), experimenter error ($n = 1$), and scoring one standard deviation below the mean on the Recalling Sentences subtest ($n = 2$).

2.2 Stimuli

Participants heard 120 sentences which each contained four monosyllabic words (Stelmachowicz *et al.*, 2000). Half of the sentences were high-predictability (e.g., "Pour me more tea.") and half were low-predictability (e.g., "Most birds knock tea."). The sentences were recorded by two female native speakers of American English (one of whom was the first author) and mixed with signal-dependent noise (Benkí, 2003) at −5 dB signal-to-noise ratio using a MATLAB script (Felty, 2007) [see Case *et al.* (2018) for more information on the recording procedure and acoustics of the two speakers]. To become familiarized with the noisy stimuli, participants also heard practice sentences produced by a third female native speaker of American English, which did not contain any words from the 120 experimental sentences.

2.3 Procedure

The experiment took place over 2 days one week apart (±2 days). Participants completed baseline and post-exposure sentence recognition tasks and a series of implicit voice-learning tasks that were pre-recorded and presented to participants over headphones. The implicit voice-learning tasks were standardized tests commonly used in clinical settings and were selected because they include a high quantity of speaking by the person administering the test, thus allowing participants ample opportunity to

become familiar with the talker's voice. Participants were randomly assigned to hear one of the two talkers (JEC and SSL) during this exposure portion (15 participants in the JEC condition and 16 in the SSL condition). All experimental tasks were run on a Panasonic Toughbook CF-52 laptop using E-Prime 2.0 Professional (Schneider *et al.*, 2007). Stimuli were presented binaurally over Sennheiser HD-280 headphones. A different trained research assistant who was not one of the recorded talkers administered all tasks.

Baseline and post-exposure sentence recognition. At the beginning of day 1, participants completed a self-paced sentence recognition task in which they heard 30 low-predictability sentences and 30 high-predictability sentences in random order, evenly divided across the two talkers. Before each trial, the name of the talker appeared on the screen. Participants were asked to repeat the sentence out loud. At the end of day 2, the remaining 60 sentences (30 high-predictability, 30 low-predictability) were presented to the participants, with half produced by each talker. Thus, there was no overlap between the sentences heard at baseline and at post-learning.

Implicit voice-learning tasks. The same implicit voice-learning tasks used for the adults in Case *et al.* (2018) were used in this experiment. The crucial difference was that all spoken components of the tasks, including the directions, were recorded by the two talkers (JEC and SSL), rather than being administered by the two talkers in person. Importantly, the same two talkers served as the familiar voices in both Case *et al.* (2018) and the current study. Recordings for the implicit voice-learning stimuli were made with the same equipment as the recordings for the sentence recognition task. Participants assigned to JEC's group were presented with voice-learning tasks spoken by JEC, while those in the SSL group listened to SSL. Implicit voice-learning tasks were administered without the addition of background noise.

On Day 1, after the baseline spoken sentence recognition task, participants completed three subtests of the *Clinical Evaluation of Language Fundamentals—4th edition* (*CELF-4*): *Recalling Sentences*, *Word Definitions*, and *Understanding Spoken Paragraphs* (Semel *et al.*, 2003). On Day 2, they completed *Understanding Spoken Paragraphs* a second time and also completed the *Semantic Relationships* subtest of the *CELF-4*. All *CELF-4* subtests were presented in the voice of one of the two talkers (JEC, SSL). The Understanding Spoken Paragraphs subtest was administered a second time on Day 2 to make up for the lack of the verbally presented participant questionnaire and for the lack of voice familiarization with the hearing screening that was used in Case *et al.* (2018). Similar to our previous study, participants received approximately 55 min of exposure to the talker's voice, comprised of 35 min on Day 1 and 20 min on Day 2.

2.4 Analysis

Each verbal response was transcribed live during the experiment and was also transcribed independently by a research assistant based on the recorded audio. Any discrepancies between the two transcriptions were resolved by a third coder. The response to each sentence was coded as a correct or incorrect sentence identification, using the same procedure as in Case *et al.* (2018) (adapted from Stelmachowicz *et al.*, 2000). For a participant's response to be coded as correct, the response only had to include all four words from the original sentence, even if they were produced in the wrong order, if extra words were inserted, or if any of the words had different inflectional suffixes than the original.

Accuracy was modeled with a logistic mixed-effects regression fit with lme4 (Bates *et al.*, 2015; R Core Team, 2016), using the same model parameters as Case *et al.* (2018).[1] The model included sum-coded (−0.5 vs 0.5) parameters for Time (baseline vs post-learning), Talker Type (unfamiliar vs familiar), Talker Identity (SSL vs JEC), and Predictability (low vs high) and all interactions up to the four-way interaction. Also included were by-participant intercepts; by-participant slopes for Time, Talker Type, and their interaction; by-sentence intercepts; and by-sentence slopes for Time, Talker Type, and their interaction.

The emmeans package in R (Lenth, 2018) was used to estimate marginal means for each cell in the design and to evaluate the significance of the relevant marginal contrasts. The crucial contrast is the difference in the effect size of Time for familiar versus unfamiliar talkers, averaging over all four combinations of Talker Identity and Predictability. All differences (M) reported below are in log-odds, and all p-values are calculated based on the Wald z-statistic for the estimate of the relevant contrast, with family-wise corrections for multiple comparisons where appropriate (using Holm's method; corrected values are indicated with an asterisk as p^* below).

Because of the coding scheme, many of the marginal effect sizes reported below are equivalent to the model parameter coefficients.

3. Results

There was an overall significant improvement in accuracy from baseline to post-learning (marginal effect size $M = 0.33$ difference in log-odds, $z = 2.68$, $p = 0.007$). Crucially, the interaction with talker type was significant ($M = 0.38$, $z = 1.99$, $p = 0.046$),[2] such that improvement on sentences produced by familiar talkers was greater than on sentences produced by unfamiliar talkers. For sentences produced by familiar talkers, the improvement was significant ($M = 0.53$, $z = 3.36$, $p^* = 0.002$), while the improvement was not significant for sentences produced by unfamiliar talkers ($M = 0.14$, $z = 0.90$, $p^* = 0.367$). The left panel of Fig. 1 shows the model estimates of accuracy (transformed to the absolute probability of a correct response) for testing time and talker type, averaging across the levels of talker identity and sentence predictability. The right panel shows model estimates of accuracy for the four combinations of talker identity and predictability.

As in Case *et al.* (2018), there were additional overall effects of talker identity ($M = 0.87$, $z = 8.00$, $p < 0.001$), where participants were more accurate when listening to sentences produced by JEC; and of predictability ($M = 2.3$, $z = 8.79$, $p < 0.001$), with higher accuracy for high-predictability versus low-predictability sentences. There was also an additional significant two-way interaction between talker identity and predictability ($M = 0.42$, $z = 2.06$, $p = 0.039$), such that there was a larger effect of predictability for sentences produced by SSL. The model parameter estimates (fixed-effects estimates $\hat{\beta}$ and standard deviations of random-effects estimates $\hat{\sigma}$) are shown in Table 1.

4. Discussion and conclusion

We found that short-term implicit voice-learning generated a Familiar Talker Advantage in an experiment in which the encoding (learning) and retrieval (test) phases were conducted in similar conditions. While our previous work failed to find spoken language processing benefits following short-term voice exposure (Case *et al.*, 2018), that study used different conditions in each phase. Specifically, encoding occurred during live in-person interactions, and testing occurred with pre-recorded stimuli over headphones. In contrast, the current study used pre-recorded stimuli over headphones in both the encoding and retrieval phases and we found that spoken language processing did indeed improve more when participants listened to a familiar as compared to an unfamiliar talker. Because both studies involved short-term implicit voice exposure, it is unlikely that the failure of our previous study to find the Familiar Talker Advantage was simply due to inadequate learning of voice information.

In cases of short-term exposure, it is possible that a Familiar Talker Advantage only arises when encoding and retrieval conditions are sufficiently similar. To test this possibility, we pooled the data from the current study (with similar encoding-retrieval

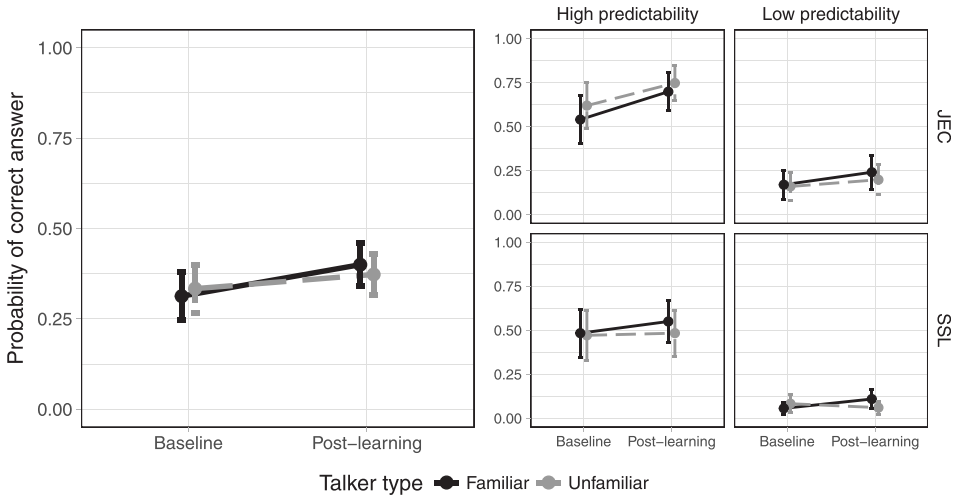

Fig. 1. Mean probability of correct sentence identification (y-axis) by testing time (x axis), estimated from the model. Error bars show ± the standard error of the model-estimated mean probabilities. The left panel shows marginal means averaged over the levels of predictability and talker identity, while the right panel shows the marginal means for each combination of levels for all four predictors.

Table 1. Parameter estimates (in log-odds) for the logistic mixed-effects model. Columns show the parameter names (column 1), the fixed-effect parameter estimates (column 2), and the standard deviation for the random-effects parameter estimates (columns 3–4). For the fixed-effects, standard errors are given in parentheses and stars indicate significance based on the Wald z-statistic (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

Parameter	$\hat{\beta}$ (SE)	$\hat{\sigma}$ for by-participant parameter estimates	$\hat{\sigma}$ for by-sentence parameter estimates
Intercept	−0.83 (0.15)***	0.45	1.31
Time	0.33 (0.12)**	0.27	0.51
Talker Type	0.04 (0.10)	0.20	0.09
Talker Identity	0.87 (0.11)***	—	—
Predictability	2.3 (0.26)***	—	—
Time × Talker Type	0.38 (0.19)*	0.15	0.31
Time × Talker Identity	0.32 (0.21)	—	—
Talker Type × Talker Identity	−0.20 (0.37)	—	—
Time × Predictability	0.13 (0.24)	—	—
Talker Type × Predictability	−0.20 (0.19)	—	—
Talker Identity × Predictability	−0.42 (0.2)*	—	—
Time × Talker Type × Talker Identity	−0.51 (0.41)	—	—
Time × Talker Type × Predictability	−0.47 (0.39)	—	—
Time × Talker Identity × Predictability	0.32 (0.41)	—	—
Talker Type × Talker Identity × Predictability	−0.49 (0.36)	—	—
Time × Talker Type × Talker Identity × Predictability	−0.74 (0.72)	—	—

conditions) with the data from the adult participants in Case *et al.* (2018) (with different encoding-retrieval conditions), and fit an exploratory model following the procedure in Sec. 2.4, with additional fixed-effects parameters for Encoding Type (similar vs different) and its interaction with Time and Talker Type. A significant three-way interaction would suggest that the Familiar Talker Advantage is moderated by encoding specificity (i.e., whether the encoding and retrieval conditions are similar). However, the interaction was non-significant ($p = 0.619$). As a second test, we evaluated whether there was evidence for an overall Familiar Talker Advantage, averaging across both experiments. This contrast was also non-significant, though in the expected direction ($M = 0.22$, $z = 1.64$, $p = 0.102$). Thus, while the current study provides good evidence that a Familiar Talker Advantage arises *at least* when encoding and retrieval conditions are similar, the evidence is equivocal as to whether or not the Familiar Talker Advantage is necessarily limited by encoding specificity.

More generally, our finding contributes to existing knowledge on the Familiar Talker Advantage by crucially demonstrating that short-term implicit voice-learning is sufficient to lead to the advantage [see Souza *et al.* (2013) and Johnsrude *et al.* (2013) on *long-term* implicit learning]. This finding may have clinical implications such as how well a client performs on spoken language processing tests with either a speech-language pathologist or audiologist whose voice may become familiar as a result of repeated exposure through evaluation and treatment. Previous work has explored the limits of the Familiar Talker Advantage, such as whether it is moderated by learning ability (Levi *et al.*, 2011; Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994) or sentence versus single-word recognition (Nygaard and Pisoni, 1998), and the current study explores the role of encoding specificity. Taken together, this line of research points to the need to better understand the context-dependence of learned advantages in spoken language processing (e.g., Reinisch *et al.*, 2014).

Acknowledgments

This work was supported in part by a grant from the NIH-NIDCD, Grant No. 1R03DC009851 (S.V.L.). We would like to thank Gabrielle Alfano, Stephanie Lee, Maddy Lippman, Rebecca Piper, and Ashley Quinto for help with data collection and the children and families for their participation. Portions of this work were presented at the American Speech Language and Hearing Annual Convention (2016) and at the Symposium on Research in Child Language Disorders (2017).

References and links

[1]Following Case *et al.* (2018), we also conducted a planned analysis of individual word accuracy, in addition to the planned analysis of sentence accuracy described in the main text. The crucial interaction was non-significant for the word-level analysis ($z = 0.94$, $p = 0.34$), though in the predicted direction ($M = 0.12$).

[2]This analysis was planned using the same procedure as in Case *et al.* (2018), which had reported a null result. A reviewer asks if this effect is significant if a likelihood ratio test is used instead of a Wald z test; it is not [$\chi^2(1) = 3.75$, $p = 0.052$].

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**(1), 1–48.

Benkí, J. R. (**2003**). "Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition," J. Acoust. Soc. Am. **113**(3), 1689–1705.

Bent, T., Buchwald, A., and Pisoni, D. B. (**2009**). "Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech," J. Acoust. Soc. Am. **126**(5), 2660–2669.

Bradlow, A. R., and Bent, T. (**2008**). "Perceptual adaptation to non-native speech," Cognition **106**(2), 707–729.

Case, J., Seyfarth, S., and Levi, S. V. (**2018**). "Does implicit voice learning improve spoken language processing? Implications for clinical practice," J. Speech, Lang., Hear. Res. **61**(5), 1251–1260.

Felty, R. A. (**2007**). "Context effects in spoken word recognition of English and German by native and non-native listeners," Ph.D. thesis, Michigan State University.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004.

Kreitewolf, J., Mathias, S. R., and von Kriegstein, K. (**2017**). "Implicit talker training improves comprehension of auditory speech in noise," Frontiers Psychol. **8**, 1584.

Lenth, R. (**2018**). "Emmeans: Estimated marginal means, aka least-squares means," R Package Version 1.2.1, https://cran.r-project.org/package=emmeans (Last viewed October 22, 2018).

Levi, S. V. (**2015**). "Talker familiarity and spoken word recognition in school-age children," J. Child Lang. **42**(4), 843–872.

Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2011**). "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?," J. Acoust. Soc. Am. **130**(6), 4053–4062.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(1), 42–46.

R Core Team. (**2016**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org/ (Last viewed October 22, 2018).

Reinisch, E., Wozny, D. R., Mitterer, H., and Holt, L. L. (**2014**). "Phonetic category recalibration: What are the categories?," J. Phonetics **45**, 91–105.

Schneider, W., Eschman, A., and Zuccoloto, A. (**2007**). E-prime 2.0 Professional, Psychological Software Tools, Inc., Pittsburgh, PA.

Semel, E. M., Wiig, E. H., and Secord, W. (**2003**). *Clinical Evaluation of Language Fundamentals, 4th ed. (CELF-4)*, The Psychological Corporation/A Harcourt Assessment Company, Toronto, Canada.

Souza, P., Gehani, N., Wright, R., and McCloy, D. (**2013**). "The advantage of knowing the talker," J. Am. Acad. Audiol. **24**(8), 689–700.

Stelmachowicz, P. G., Hoover, B. M., Lewis, D. E., Kortekaas, R. W., and Pittman, A. L. (**2000**). "The relation between stimulus context, speech audibility, and perception for normal-hearing and hearing-impaired children," J. Speech, Lang., Hear. Res. **43**(4), 902–914.

Tulving, E., and Thomson, D. M. (**1973**). "Encoding specificity and retrieval processes in episodic memory," Psychol. Rev. **80**(5), 352–373.

Van Engen, K. J. (**2012**). "Speech-in-speech recognition: A training study," Lang. Cognitive Process. **27**(7–8), 1089–1107.

Yonan, C. A., and Sommers, M. S. (**2000**). "The effects of talker familiarity on spoken word identification in younger and older listeners," Psychol. Aging **15**(1), 88–99.