https://doi.org/10.3758/s13414-023-02790-6

Attention modulates perceptual learning of non-native-accented speech

Christina Y. Tzeng[1] · Marissa L. Russell[2] · Lynne C. Nygaard[3]

Accepted: 11 September 2023 / Published online: 23 October 2023

Abstract

Listeners readily adapt to variation in non-native-accented speech, learning to disambiguate between talker-specific and accent-based variation. We asked (1) which linguistic and indexical features of the spoken utterance are relevant for this learning to occur and (2) whether task-driven attention to these features affects the extent to which learning generalizes to novel utterances and voices. In two experiments, listeners heard English sentences (Experiment 1) or words (Experiment 2) produced by Spanish-accented talkers during an exposure phase. Listeners' attention was directed to lexical content (transcription), indexical cues (talker identification), or both (transcription + talker identification). In Experiment 1, listeners' test transcription of novel English sentences spoken by Spanish-accented talkers showed generalized perceptual learning to previously unheard voices and utterances for all training conditions. In Experiment 2, generalized learning occurred only in the transcription + talker identification condition, suggesting that attention to both linguistic and indexical cues optimizes listeners' ability to distinguish between individual talker- and group-based variation, especially with the reduced availability of sentence-length prosodic information. Collectively, these findings highlight the role of attentional processes in the encoding of speech input and underscore the interdependency of indexical and lexical characteristics in spoken language processing.

Keywords Perceptual learning · Attention · Spoken language perception · Non-native-accented speech

Introduction

Speech perception is a complex process, requiring listeners to extract a meaningful message from a highly variable, multi-layered signal. Despite variation due to factors such as phonetic context and speaking rate, as well as talker-specific characteristics, such as age, gender, language background, and idiosyncratic differences in pronunciation across individual talkers, listeners typically achieve stable perception. Even in listening situations characterized by increased perceptual difficulty due to acoustic-phonetic deviations, as when faced with non-native-accented speech (Munro & Derwing, 1995), listeners readily adapt with experience (e.g., Clarke & Garrett, 2004; Bradlow & Bent, 2008; Tzeng et al., 2016; Xie & Myers, 2017). A now extensive literature has documented evidence for listeners' ability to adapt to non-standard pronunciations. What remains unclear are the cognitive mechanisms that enable listeners to understand voices that they have not encountered before. A central goal of the current work was to characterize whether attentional modulation during exposure to variation in speech affects listeners' likelihood of achieving adaptation.

Founded on the assumption that variation due to surface characteristics impedes spoken word recognition, traditional theoretical views on speech perception have maintained that perceptual constancy is achieved through a normalization process that discards talker-related variability (e.g., Ladefoged & Broadbent, 1957; Magnuson & Nusbaum, 2007; Mullennix et al., 1989). According to normalization accounts, the perceptual system converts the speech signal to a canonical form, which can then be more efficiently aligned with existing representations stored in memory. However, these views have been challenged by accumulating evidence that listeners do retain and use talker-dependent variability in representations of spoken utterances during spoken word

✉ Christina Y. Tzeng
christina.tzeng@sjsu.edu

[1] Department of Psychology, San José State University, 1 Washington Sq, San José, CA 95192, USA

[2] Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, USA

[3] Department of Psychology, Emory University, Atlanta, GA, USA

recognition (Goldinger, 1996, 1998; Palmeri et al., 1993). To accommodate findings that familiarity with talker variation facilitates spoken language processing (e.g., Clarke & Garrett, 2004; Nygaard et al., 1994), more recent models acknowledge listeners' sensitivity to not only the distribution of variability (e.g., Clayardset al., 2008; Toscano & McMurray, 2012), but also to the systematic co-variation between acoustic-phonetic variation and socio-indexical variables (e.g., individual talker identity, talker's inferred regional origin; Kleinschmidt & Jaeger, 2015; Sumner et al., 2014). Under these types of accounts, listeners dynamically infer higher-order categories (e.g., group membership of talkers) using the current input alongside representations of previously encountered variation in speech sounds.

Non-native-accented speech offers a unique window for examining the processes by which listeners overcome variability in the speech signal. Not only is non-native-accented speech characterized by idiosyncratic differences in pronunciation across individual talkers, but it is also marked by systematic acoustic-phonetic deviations from the sound categories of native speakers. In order to achieve comprehension that is talker-independent, listeners must learn to portion talker-specific variation from variation attributed to regularities shared among speakers of the same non-native accent. Abundant evidence suggests that listeners can achieve talker-independent learning of non-native-accented speech (e.g., Baese-Berk et al., 2013; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Sidaras et al., 2009; Tzeng et al., 2016; Witteman et al., 2013; Xie et al., 2018, 2021).

Sidaras et al. (2009), for example, found that transcription accuracy of novel tokens spoken by both familiar and unfamiliar Spanish-accented talkers at test was reliably more accurate for listeners who transcribed utterances spoken by Spanish-accented talkers during an exposure phase than those who did not. This suggests that listeners can, with experience, learn the systematic acoustic-phonetic regularities of a non-native accent such that they can understand previously unheard talkers and utterances from the same accent group. After exposure to utterances spoken by a native Mandarin talker in a cross-modal word-matching task, listeners in Xie et al. (2018) showed both faster and more accurate processing at test of utterances produced by the same talker, as well as by a different talker of the same accent, suggesting that exposure to only a few minutes of non-native-accented speech attenuates listeners' initial processing difficulty with non-standard pronunciations. Taken together, these findings suggest that listeners can meaningfully and rapidly restructure processing of variation to enhance the recognition of novel utterances spoken by unfamiliar talkers of the same accent to which they have not been directly exposed.

In order for learning to successfully generalize to novel utterances, listeners may not only require exposure to relevant stimuli, but they may also need to perform a perceptual task that entails engagement with the presented tokens (e.g., Borrie et al., 2013; Drouin & Theodore, 2022; Wright et al., 2015). Wright et al. (2015), for instance, assessed the extent to which passive exposure to versus active engagement with non-native-accented speech yielded robust perceptual learning. Relative to participants who passively listened to Mandarin-accented sentences during an exposure phase, those who transcribed sentences showed reliably higher levels of transcription accuracy for novel sentences spoken by an unfamiliar Mandarin-accented talker at test. Assessing the role of explicit attention in the perceptual learning of noise-vocoded speech, a spectrally degraded signal that induces processing costs for listeners, Huyck and Johnsrude (2012) found that listeners who attended to noise-vocoded sentences and reported what they heard during exposure demonstrated more accurate subsequent novel sentence comprehension than those who attended to either simultaneously presented auditory or visual distractors during the exposure phase. These findings provide support for the notion that explicit attention to the speech signal during learning enhances listeners' ability to adapt to the perceptual consequences of systematic variation.

Allocation of attentional resources has also been found to impact other perceptual and cognitive processes implicated in spoken language perception. Examining the effects of listener attention on speech segmentation, Toro et al. (2005) found that whereas attention directed toward the target speech stream facilitated listeners' word segmentation performance, diverting listeners' attention to a simultaneously presented auditory or visual stream negatively impacted word extraction ability. When attentional resources are directed away from task-relevant cues, here the recurring transitional regularities in the continuous speech stream, target task performance is impaired.

A critical question that emerges from the above-described findings is whether for a learning outcome that requires linguistic processing (e.g., word recognition), attention needs to be specifically directed toward linguistic properties (e.g., acoustic-phonetic components of spoken lexical items) during exposure. Seitz et al. (2010) examined perceptual learning of speech formant transitions by pairing single formant transitions that had frequency sweeps at sub-threshold detection levels with animal sounds. In an implicit learning condition, participants were asked to attend to and make judgments about the animal sounds. Other participants heard only the single formant transitions and were asked to make explicit identification judgments. In a subsequent formant discrimination task, only participants in the implicit learning condition exhibited improved formant discrimination performance, suggesting that for formant discrimination, attention directed explicitly to the to-be-learned acoustic components of formant transitions, a type of linguistic property, is unnecessary. These results, along with others (e.g., Vlahou et al.,

2012) raise the possibility that for listeners to learn stable speech sound category structures, the relevant properties need not be the focus of explicit task-directed attention, but rather simply correlated with other aspects of the perceptual event. Indeed, the attended-to property within the learning context need not even be linguistic at all to promote learning.

Perhaps intuitively, however, for perceptual learning of systematic variation in speech, behavioral tasks that require listeners to explicitly attend to linguistic information highlight the systematicity of acoustic-phonetic deviations from standard pronunciations. During word or sentence transcription tasks, for instance, listeners engage in lexical processing, using phonological and morphological information to facilitate word recognition. Comprehension of non-native-accented speech requires listeners to do this while also tracking systematic differences in pronunciation due to the speaker's native language background and proficiency (e.g., learning that Spanish-accented speakers of English often produce the /I/ vowel as in "sit" as /i/, as in "seat"; Alexander & Nygaard, 2019; Flege et al., 2003; Sidaras et al., 2009). For non-native-accented speech, variation due to idiolect is nested within variation that is shared among talkers of the same language background. That is, although individual non-native speakers have unique voice characteristics, the pattern of their deviations from native pronunciations due to shared characteristics of their native language phonology is systematic and relatively predictable. These nested sources of variation present a complex perceptual learning task requiring listeners to apportion the variance due to individual talkers' voices from the shared variation due to shared accented characteristics.

When exposed to speech by multiple non-native-accented talkers, as has been the approach in several empirical investigations of talker-independent perceptual learning (e.g., Bradlow & Bent, 2008; Tzeng et al., 2016; but see, Xie et al., 2021), listeners do learn to distinguish the variability attributed to a particular language group (e.g., Spanish-accented speakers) from that attributed to any individual talker. With knowledge of the structure of individual talker versus accent-based variation, listeners can adjust their representations of target speech sounds in a manner that facilitates the generalization of acoustic-phonetic adaptation from one talker to other talkers of the same non-native accent. Implicated in this characterization of non-native-accented speech perception is the possibility that directing attention toward *indexical cues*, or cues that signal talker identity and group membership (i.e., native language community), might uniquely facilitate listeners' extraction of an accent or group-based schema for speech sound categories.

Evidence to support this possibility comes from empirical investigations of listeners' ability to adapt to other types of non-standard speech (Davis et al., 2005; Dorman et al., 1997). Loebach et al. (2008), for example, examined adaptation to noise-vocoded speech. As indexed by a transcription task at test with novel noise-vocoded sentences, listeners who completed either a talker-identification or a transcription task during an exposure phase reliably outperformed those who completed a gender identification task. Importantly, the magnitude of the facilitative effect of exposure was comparable for the groups who completed the talker-identification and transcription tasks, suggesting that attention to either linguistic or indexical details constrains the acoustic-phonetic space for listeners, yielding increased levels of intelligibility. Examining perceptual adaptation to dysarthric speech, Borrie et al. (2013) found that listeners who completed a word-identification or a speaker-identification task during training demonstrated similar and higher intelligibility gains on a novel phrase transcription task presented at test, relative to listeners who were passively exposed to training stimuli. Although this study did not address the extent to which learning generalizes to novel talkers, the findings suggest that attention directed toward indexical, as well as linguistic, properties of speech yielded learning that extended to phonetic contexts that listeners had not heard during exposure.

Accounts of speech perception are largely agnostic with respect to the role of attention in listeners' ability to overcome variation. Hypothetically, listeners may, by attending to particularly informative or relevant details in the speech stream, be able to form or modify linguistic processing and representation to infer speech sound categories that generalize to novel utterances produced by previously unheard non-native-accented talkers. However, there is little empirical evidence for whether and how listeners do this. In the present work, we tested the hypothesis that listeners' ability to generalize learning to novel non-native-accented talkers and utterances would vary as a function of which aspects of the utterance (linguistic vs. indexical) their attention was directed toward during exposure. In two experiments, we held the exposure tokens (sentences in Experiment 1; words in Experiment 2) constant across conditions, varying only the behavioral task. Our primary goal was to assess the impact of listeners' attentional focus on the likelihood of generalized learning. Although existing evidence points to a facilitative effect of active versus passive listening on the likelihood of learning previously unheard non-native-accented utterances (e.g., Wright et al., 2015), the possibility that attention to different aspects of the accented utterance could have differential effects on learning remains unexplored. Given findings suggestive of the importance of indexical cues in lexical processing (e.g., Creel et al., 2008; Dahan et al., 2008; Nygaard et al., 1994; Nygaard & Pisoni, 1998), a second goal was to examine the relative benefits of attending to linguistic versus indexical cues for separating individual talker variation from group-based (i.e., non-native accent) variation.

Table 1 Accentedness and intelligibility for Spanish-accented talkers

Speaker Group	Gender	Mean Accentedness (sentences)	Mean Intelligibility (sentences, %)	Mean Intelligibility (words, %)	Age of English Acquisition (years)	Age of Arrival in the U.S. (years)	Length of Residence in the U.S. (years)
Group 1	Female	5.6	75.6	32.9	28	27	4
	Female	3.1	89.8	68.8	2	3	34
	Male	4.8	65.9	42.9	25	32	3
	Male	2.8	90.5	60.3	20	22	7
Group 2	Female	6.2	74.6	48.9	10	10	2
	Female	4.3	85.5	35.2	27	27	15
	Male	4.8	89.0	54.4	16	27	1
	Male	3.6	81.8	49.2	10	22	15

Listeners rated the accentedness of each sentence on a 7-point Likert-type scale, from 1 = not accented to 7 = very accented

Experiment 1

Experiment 1 assessed the role of attentional focus on perceptual learning of non-native-accented speech. If listeners must actively attend to linguistic information to learn accent regularities, then relatively more robust generalization will be observed for listeners who are encouraged to attend to linguistic rather than indexical cues during exposure. If, however, generalization can also be achieved by attending to indexical cues, or if learning the structure of variation is not contingent on the distribution of attentional resources, comparable generalization will also be observed for those who attend to indexical cues. An additional possibility is that maximally robust generalization occurs when listeners explicitly attend to *both* types of variation.

Method

Participants

Ninety-four Emory University undergraduates (64 female, 30 male) received course credit for their participation. All were native monolingual speakers of American English and reported no history of speech or hearing disorders. Data from two participants were excluded due to equipment malfunction, leaving data from 92 participants included in the reported analyses.

Stimuli

Eight native speakers of Spanish (four female, four male) from Mexico City were selected from a set of 12 native Spanish speakers that were used in Sidaras et al. (2009). An additional four native speakers of American English (two male, two female) recorded the same materials to serve as stimuli for the control condition. All talkers were recruited from the Atlanta area.

Each talker produced 144 monosyllabic English words and 100 Harvard sentences (Rothauser et al., 1969). Sentences were monoclausal and ranged from six to ten words, with five key words per sentence (e.g., The *salt breeze came across* from the *sea*). All sentences were chosen from lists that are phonetically balanced to reflect the frequency of phonemes in English. Monosyllabic words were categorized as either easy or hard. Easy words were high-frequency words (M = 309.69; Kučera & Francis, 1967) with few (M = 38.32) low-frequency neighbors (e.g., *voice*, *reach*; Luce & Pisoni, 1998). Hard words were low-frequency words (M = 12.21) with many (M = 282.22) high-frequency neighbors (e.g., *kin*, *pawn*). Both easy and hard words were rated as being highly familiar (M = 6.97; on a scale of 1–7 with 1 being not familiar at all and 7 being highly familiar; Nusbaum et al., 1984). Recordings were re-digitized at a 22,050-Hz sampling rate, edited into separate files, and amplitude normalized.

To determine baseline intelligibility of the recorded stimuli, separate groups of native English-speaking listeners transcribed all 144 words and 100 sentences for each of the 12 talkers (ten listeners per non-native-accented talker). The proportion of correctly transcribed words was calculated across listeners for each of the twelve talkers. An additional ten listeners rated the accentedness of ten sentence-length utterances from each of the 12 talkers. Listeners rated the accentedness of each sentence on a 7-point Likert-type scale, from 1 = not accented to 7 = very accented. Sentences were presented in the clear for intelligibility transcription and accentedness rating measures. Table 1 presents detailed demographic information, along with the mean accentedness and intelligibility ratings, for the eight selected non-native-accented talkers (see also Sidaras et al., 2009).

Fifty-six Harvard sentences (Experiment 1) and 104 monosyllabic words (Experiment 2) were selected to use as stimuli in the current study. Two groups of four talkers (two male and two female) were created as exposure and test groups. Talker groups were equated overall for sentence intelligibility and accentedness such that the two groups did not differ significantly on either factor (sentence intelligibility, $t(6) = .-34$, $p = .746$; accentedness, $t(6) = -.72$, $p = .497$). Within each group, two of the talkers (one male and one female) were characterized as high-intelligibility ($M_{Group\ 1} = 90.15$; $M_{Group\ 2} = 87.25$), and two as low-intelligibility talkers ($M_{Group\ 1} = 70.75$; $M_{Group\ 2} = 78.20$). Across both talker groups, intelligibility ratings were reliably higher for high-intelligibility versus low-intelligibility talkers, $t(6) = 4.12$, $p = .006$, $M_{high} = 88.70$; $M_{low} = 74.48$.

Procedure

Participants were randomly assigned to one of four conditions, each of which included an exposure phase immediately followed by a test phase. Listeners completed the experiment on Dell Optiplex desktop computers using E-prime 2.0 (Schneider et al., 2002). Auditory stimuli were presented binaurally over Beyerdynamic DT100 headphones at approximately 65 dB SPL.

Exposure phase All participants heard 36 Harvard sentences spoken four times each, once by each of the four talkers. In three of the four conditions, sentences were spoken by four Spanish-accented talkers such that participants heard the same Spanish-accented tokens across the transcription, talker ID, and transcription + talker ID conditions. However, the participant's task differed across conditions. In the transcription condition, participants transcribed each sentence they heard. In the talker ID condition, participants identified the talker as one of four individuals (Alice, Bob, Carol, and Mike) by pressing the appropriately labeled buttons on a keyboard. Participants in the transcription + talker ID condition transcribed sentences for half of the exposure trials and identified talkers in the other half, with task blocked such that participants alternated between transcription and talker identification every 12 trials. The fourth condition served as a control condition during which participants heard and transcribed the same 36 sentences spoken by four native English-speaking talkers.

After each response, participants in all conditions received corrective feedback tailored to each task (either the target sentence or the talker's name presented on the computer screen). Exposure trials were pseudo-randomized in blocks such that participants heard all 36 sentences produced by one of the four talkers in each block. Sentence repetitions (with different talker-sentence pairings) occurred across the four blocks such that participants heard each of the 36 sentences four times in the exposure phase. For the experimental conditions, speaker group was counterbalanced across exposure and test phases such that half the listeners in each condition heard group 1 during training (and group 2 at test), and half heard group 2 during training. For the control condition, half of the listeners heard group 1 at test, and half heard group 2.

Generalization test phase At test, participants in all conditions heard 20 novel sentences spoken by four Spanish-accented talkers that they did not hear during exposure. Test sentences were mixed in white noise (+10 dB signal-to-noise ratio), with sentence-speaker pairings randomized. Whereas all training stimuli were presented without noise to maximize the clarity of the talkers' utterances, mixing the test sentences in white noise served to increase task difficulty and introduce variability in participants' test performance (e.g., Tzeng et al., 2016). Participants transcribed five sentences spoken by each of the four talkers without corrective feedback. Both the exposure and test phases were self-paced and together took approximately 30 minutes to complete.

Analysis

For all analyses, trial-level responses were fit to mixed-effects models using the *lme4* package (Bates et al., 2015) in R (version 3.6.0; R Development Core Team, 2019), with *p*-values for mixed-effects analyses obtained using the *lmerTest* package (Kuznetsova et al., 2017). All fixed effects were dummy coded. Random-effect structure was the maximal structure that would allow model convergence. Best-fitting models were determined using additive stepwise model comparisons using log-likelihood ratio tests (Baayen et al., 2008).[1]

Results

Exposure

Figure 1 shows the proportion of correct responses for each sentence repetition. For the talker-identification task in the talker ID and transcription + talker ID conditions, each response was coded as either correct (1) or incorrect (0). For the transcription task in the control, transcription, and transcription + talker ID conditions, the proportion of key words correct was calculated for each sentence, with

[1] Speaker group was included as a random effect in model comparisons for all statistical analyses. To avoid model overfitting and model non-convergence, speaker group was included in the reported models as a random effect only when its addition to the models accounted for significantly more variance than when it was not included.

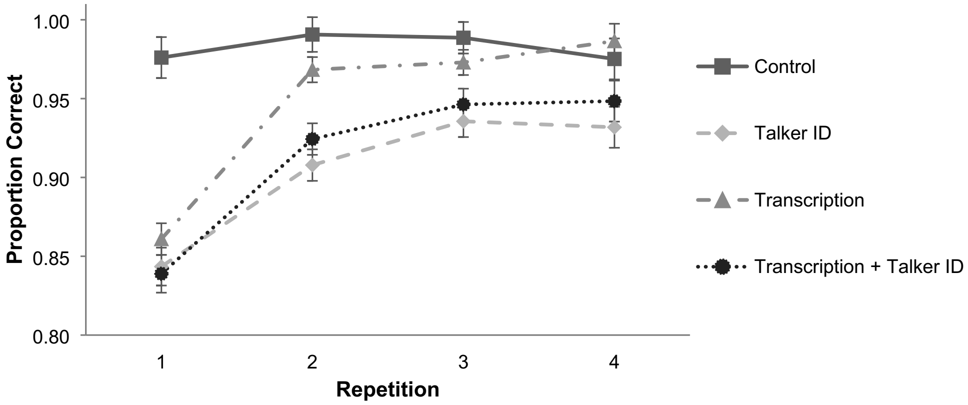

Fig. 1 Task performance for each sentence repetition during training in Experiment 1. Error bars represent standard error of the mean. Across the three experimental conditions, participants completed different tasks upon hearing the same training stimuli. In the talker ID condition, participants (n = 22) completed a four-alternative forced-choice talker-identification task. In the transcription condition, participants (n = 30) transcribed each sentence they heard. Participants in the transcription + talker ID (n = 21) condition transcribed sentences for half of the exposure trials and identified talkers in the other half, with task blocked such that participants alternated between transcription and talker identification. Listeners in the control condition (n = 19) transcribed the same sentences presented in the experimental conditions but spoken instead by four native American English-speaking talkers rather than non-native English-speaking talkers. Note because performance was high overall, the y-axis is truncated

homophones (e.g., two, too), regular verb tense changes (e.g., seems, seemed), and clearly identifiable words that contained minor typographical errors (e.g., chiken, squirel) coded as correct. Proportion correct for the transcription + talker ID condition reflects the average of performance on the two task types.

As the measurement of task accuracy differed across conditions, the trajectory of training performance was assessed separately for each condition. Four mixed-effects models assessed the extent to which training performance varied as a function of sentence repetition for each condition, fitting random intercepts by participant and sentence. Including repetition in the model as a fixed effect reliably improved model fit over a model that included just random effects for all conditions, χ^2_{Control} (3) = 27.15, $p < .001$; $\chi^2_{\text{Talker ID}}$ (3) = 52.40, $p < .001$; $\chi^2_{\text{Transcription}}$ (3) = 619.61, $p < .001$; $\chi^2_{\text{Transcription + Talker ID}}$ (3) = 104.58, $p < .001$.

Pairwise comparisons assessing the effect of repetition for each condition were run using the *emmeans* package (Lenth, 2018), with p-values adjusted for multiple comparisons using the Tukey method. Training task performance improved reliably between repetitions 1 and 2 for all conditions, $\beta_{\text{Control}} = -.015$, $SE = .004$, $p < .001$; $\beta_{\text{Talker ID}} = -.064$, $SE = .01$, $p < .001$; $\beta_{\text{Transcription}} = -.107$, $SE = .006$, $p < .001$; $\beta_{\text{Transcription + Talker ID}} = -.085$, $SE = .01$, $p < .001$. In addition, in the control condition, response accuracy reliably decreased between repetitions 3 and 4 ($\beta = -.013$, $SE = .004$, $p = .003$) and in the transcription condition, response accuracy increased marginally between repetitions 3 and 4 ($\beta = .014$, $SE = .006$, $p = .067$). Response accuracy was reliably higher at repetition 4 than at repetition 1 for all experimental conditions ($\beta_{\text{Talker ID}} = -.089$, $SE = .01$, $p < .001$; $\beta_{\text{Transcription}} = -.126$, $SE = .006$, $p < .001$, $\beta_{\text{Transcription + Talker ID}} = -.110$, $SE = .01$, $p < .001$), suggesting that participants improved in their ability to identify the individual talkers' voices (talker ID), recognize accented lexical items (transcription), or both (transcription + talker ID).

Generalization test

Figure 2 shows transcription accuracy (proportion key words correct) at test as a function of condition. A linear mixed-effects model assessed the extent to which transcription accuracy varied across conditions, fitting random intercepts by participant and sentence. Including condition in the model as a fixed effect reliably improved model fit over a model that included only random effects, χ^2 (3) = 22.76, $p < .001$, suggesting that test performance reliably differed across conditions. Pairwise comparisons assessing the effect of condition were run on the main model using the *emmeans* package (Lenth, 2018), with p-values adjusted for multiple comparisons using the Tukey method. Transcription accuracy in the talker ID ($\beta = -.13$, $SE = .03$, $p < .001$), transcription ($\beta = -.09$, $SE = .02$, $p = .002$), and transcription + talker ID ($\beta = -.10$, $SE = .03$, $p = .002$) conditions was reliably higher than in the control condition. No other comparisons across conditions were significant at the $p = .05$ level.

Discussion

Experiment 1 assessed the extent to which listeners' attention to different aspects of the accented utterance could have differential effects on robustness of perceptual learning.

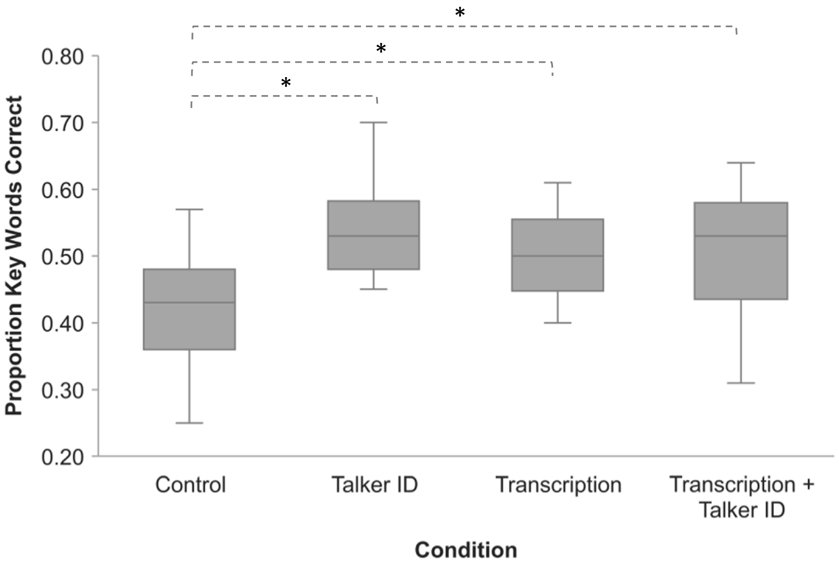

Fig. 2 Transcription accuracy at test as a function of condition in Experiment 1. Asterisks represent significance at the $p < .05$ level

Listeners in all three training conditions (transcription, talker ID, and transcription + talker ID) exhibited reliably better transcription accuracy of unfamiliar talkers and utterances than in the control condition, during which listeners transcribed sentences spoken by native English-speaking talkers in the exposure phase. Test performance did not differ across the transcription, talker ID, and transcription + talker ID conditions suggesting that attention to *either* linguistic or indexical details yields increased intelligibility for non-native-accented speech. This result is consistent with findings (e.g., Borrie et al., 2013; Loebach et al., 2008) showing comparable levels of perceptual learning between listeners who, during exposure, completed either talker-identification or transcription tasks. As the test stimuli consisted of previously unheard sentences and talkers, the facilitative effect of transcription training cannot be attributed to the learning of specific lexical items. The magnitude of perceptual benefit at test for listeners in the talker ID condition is especially notable given that unlike in the transcription condition, the task at test was entirely different to the one performed during the exposure phase. Thus, learning was not simply due to similarities between the task encountered during exposure and test. Rather, the findings from Experiment 1 suggest that attention to either linguistic or indexical details during learning facilitated word recognition and sentence processing.

A key question that follows is the extent to which these observed attentional effects on perceptual learning might vary as a function of level of linguistic processing. In addition to providing listeners with semantic and syntactic constraints, sentence-length utterances provide access to both the supra-segmental cues of an individual's voice, such as intonational contour and rhythmic patterns, as well as fine-grained acoustic-phonetic details of individual words (Baese-Berk et al., 2013). Isolated words, on the other hand, are characterized by a relatively reduced availability of supra-segmental information and require heightened levels of attention to specific acoustic-phonetic details to achieve comprehension (Greenspan et al., 1988; Nygaard & Pisoni, 1998; Sidaras et al., 2009). One possible explanation for the similar performance gains across all three training conditions in Experiment 1 is that the availability of supra-segmental cues, along with semantic and syntactic cues, in the sentence-length utterances presented during training provided a sufficient boost for comprehension regardless of which aspect (lexical vs. indexical) of the utterance listeners attended to. In Experiment 2, we presented listeners with word-length rather than sentence-length utterances to increase the processing difficulty of the exposure and test stimuli and to encourage learning of the non-native-accented variation without the prosodic cues found in sentences. The objective here was to assess the role of attention in the perceptual learning of non-native-accented speech in a more perceptually difficult task where the robustness of learning is potentially more susceptible to effects of attentional focus.

Experiment 2

In Experiment 2, listeners heard non-native-accented single words rather than sentences during exposure. We also systematically manipulated the lexical difficulty of words. Easy words were high-frequency words with few low-frequency neighbors (e.g., *voice*, *reach*), and hard words were low-frequency words with many high-frequency neighbors (e.g., *kin*, *pawn*). Previous work (e.g., Loebach et al., 2008) found that, in line with our Experiment 1 findings, transcription of novel sentences at test was equally robust for listeners who completed either a talker-identification or a transcription task during exposure. Given the increased perceptual and lexical difficulty of the current task during which participants transcribed words rather than sentences, we predicted that facilitative effects of the talker-identification and transcription tasks during exposure might be additive. That is, we expected that the most robust levels of generalization would occur in the transcription + talker ID condition when listeners explicitly attend to *both* indexical and lexical cues. Furthermore, we expected that the facilitative effect of transcription + talker ID exposure would be especially robust for easy words, as hard words might be too perceptually difficult for listeners to capitalize on the availability of indexical and lexical cues.

Method

Participants

Eighty-six Emory University undergraduates (47 female, 39 male) received course credit for their participation. All were American-English speaking monolinguals and reported no history of speech or hearing disorders. Data from four participants were excluded due to either equipment malfunction (n = 2) or fluency in another language in addition to English (n = 2), leaving data from 82 participants included in the reported analyses.

Stimuli

Participants were presented with 104 (52 easy, 52 hard) of the 144 monosyllabic words produced by the same talkers who produced the sentences included in Experiment 1.

Procedure

The exposure and test phase procedures were identical to those used in Experiment 1 except that participants responded to words (72 words during exposure, 32 novel words during test) rather than sentences, with equal representation of easy and hard words in the exposure and test phases. As in Experiment 1, all participants heard each token spoken four times during exposure, once by each of the four talkers. In the transcription, talker ID, and transcription + talker ID conditions, listeners heard the same tokens produced by the same talkers. In the control condition, participants heard and transcribed the same words spoken instead by four native English-speaking talkers. Participants in all conditions received corrective feedback (either the target word or the talker's name presented on the computer screen) after each response.

Results

Exposure

Figure 3a and b shows the proportion of correct responses for each word repetition for easy (3a) and hard (3b) words. Responses were coded as either correct (1) or incorrect (0). Word transcription accuracy was coded using the same criteria as in Experiment 1. Four logistic mixed-effects models assessed the extent to which training performance varied as a function of repetition and word difficulty for each condition, fitting random intercepts by participant and word. For all conditions except talker ID, including the interaction between repetition and word difficulty in the model reliably improved model fit over a model that included only repetition as a fixed effect, $\chi^2_{\text{Control}}(4) = 18.86$, $p < .001$, $\chi^2_{\text{Transcription}}(4) = 20.07$, $p < .001$, $\chi^2_{\text{Transcription + Talker ID}}(4) = 14.39$, $p = .006$, suggesting that the trajectory of learning across word repetitions differed as a function of word difficulty for these conditions.

Pairwise comparisons assessing the effects of word difficulty and repetition for all conditions were run using the *emmeans* package on each of the above models (Lenth, 2018), with *p*-values adjusted for multiple comparisons using the Tukey method. Mixed-effects models for the control, transcription, and transcription + talker ID conditions included the interaction between repetition and word difficulty as a fixed effect, fitting random intercepts by participant and word. The model assessing training performance in the talker ID condition included the same random effects but only repetition as a fixed effect. For easy words in the control condition, response accuracy reliably increased between repetitions 1 and 2 (β = -.84, *SE* = .28, *p* = .011) and decreased between repetitions 2 and 3 (β = .74, *SE* = .28, *p* = .041). For hard words, response accuracy marginally increased between repetitions 1 and 2 (β = -.50, *SE* = .20, *p* = .061) and repetitions 2 and 3 (β = -.60, *SE* = .25, *p* = .07).

For easy words in the transcription condition, response accuracy reliably increased between repetitions 1 and 2 (β = -.70, *SE* = .13, *p* < .001). For hard words, response accuracy reliably increased between repetitions 1 and 2 (β = -.82, *SE* = .12, *p* < .001), repetitions 2 and 3 (β = -.44, *SE* = .12, *p* = .001), and repetitions 3 and 4 (β = .34, *SE* = .12, *p* = .025). For both easy and hard words in the transcription + talker ID condition, response accuracy reliably increased between repetitions 2 and 3 (β_{easy} = -.32,

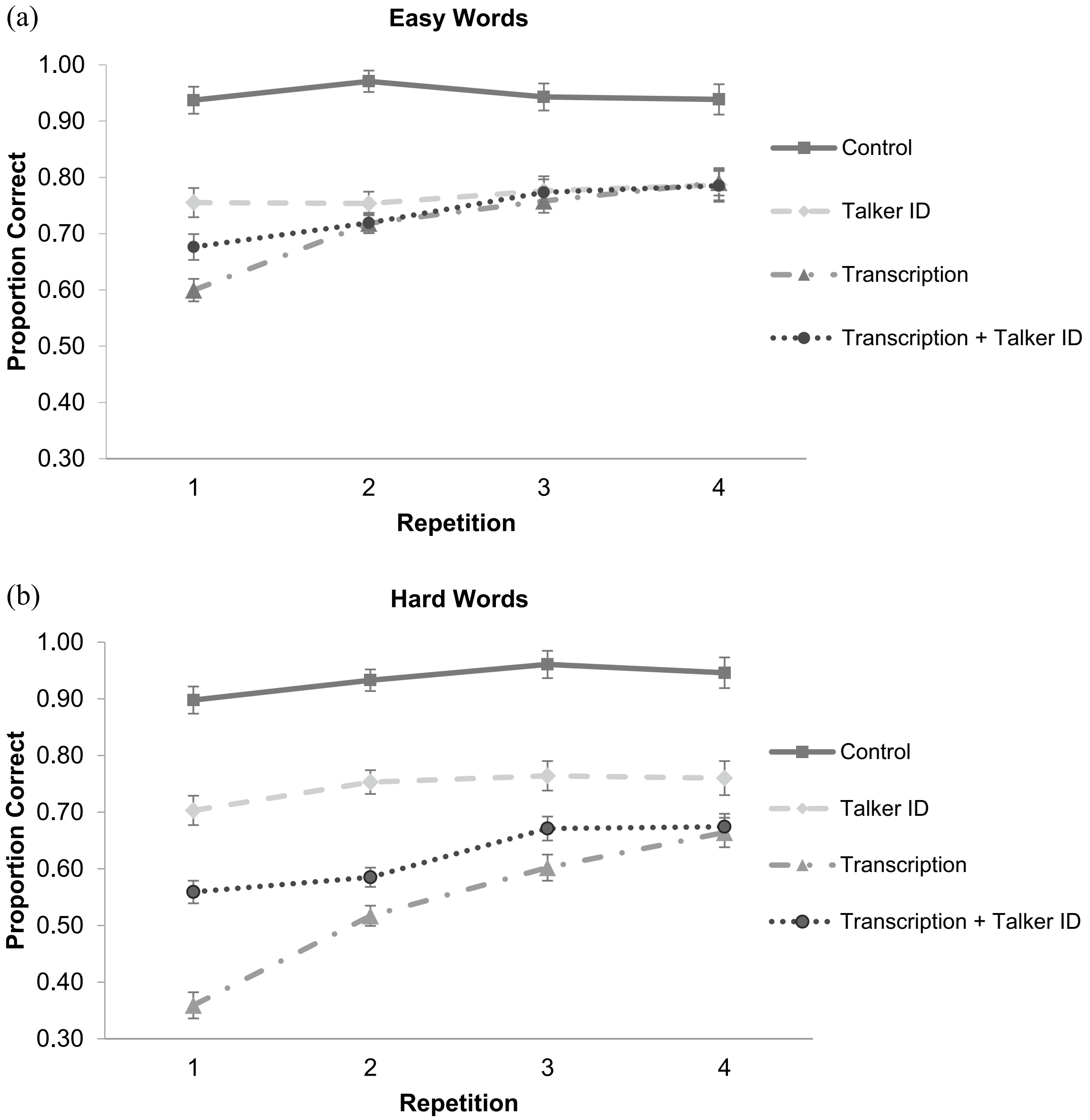

Fig. 3 Task performance for each word repetition during training in Experiment 2 for easy (**a**) and hard (**b**) words. Error bars represent standard error of the mean. In the talker ID condition, participants (n = 16) completed a four-alternative forced-choice talker-identification task. In the transcription condition, participants (n = 21) transcribed each word they heard. Participants in the transcription + talker ID condition (n = 26) transcribed words for half of the exposure trials and identified talkers in the other half, with task blocked such that participants alternated between transcription and talker identification. Listeners in the control condition (n = 19) transcribed the same words presented in the experimental conditions but spoken instead by four native American English-speaking talkers rather than non-native English-speaking talkers

SE = .11, *p* = .002, β_{hard} = -.40, *SE* = .10, *p* < .001). In the talker ID condition, although response accuracy did not reliably differ between repetitions 1 and 2, repetitions 2 and 3, or repetitions 3 and 4, response accuracy was marginally higher at repetition 4 than at repetition 1 (β = -.25, *SE* = .10, *p* = .054), suggesting that across easy and hard words, listeners did improve in their ability to identify the training talkers' voices. In the other three conditions, response accuracy was reliably higher at repetition 4 than at repetition 1 for easy words in the transcription (β = -1.21, *SE* = .13, *p* < .001) and transcription + talker ID conditions (β = -0.62, *SE* = .11, *p* < .001) and for hard words in the control (β = -0.75, *SE* = .21, *p* = .003), transcription (β = -1.6, *SE* = .12, *p* < .001), and transcription + talker ID conditions (β = -0.54, *SE* = .10, *p* < .001). Taken together, these results suggest that, overall, participants improved in their ability to identify and comprehend the exposure talkers' voices.

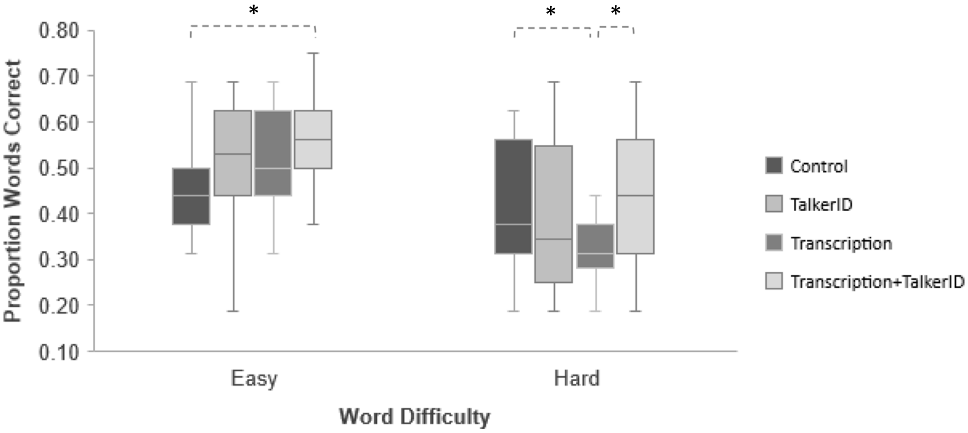

Fig. 4 Transcription accuracy at test as a function of condition and word difficulty in Experiment 2. Asterisks represent significance at the $p <$ .05 level

Generalization test

Figure 4 shows transcription accuracy (proportion words correct) at test as a function of condition and word difficulty. A logistic mixed-effects model assessed the extent to which transcription accuracy varied across condition and word difficulty (easy, hard), fitting random intercepts by participant, word, and speaker group. Including the interaction between condition and word difficulty in the model as a fixed effect reliably improved model fit over a model that included just condition and word difficulty as main effects, χ^2 (1) = 5.92, p = .015, suggesting that effect of condition on test performance varied with word difficulty.

Pairwise comparisons assessing the interaction between word difficulty and condition were run on the main model using the *emmeans* package (Lenth, 2018), with p-values adjusted for multiple comparisons using the Tukey method. For easy words, transcription accuracy in the transcription + talker ID condition was reliably higher than in the control condition (β = -.60, *SE* = .20, p = .016). For hard words, performance in the experimental conditions did not exceed that in the control condition. Instead, test accuracy in the transcription condition was reliably *lower* than in the control (β = .64, *SE* = .21; p = .011) and transcription + talker ID (β = -.65, *SE* = .19; p = .003) conditions. No other comparisons across conditions were significant at the p = .05 level.

Discussion

Whereas listeners in Experiment 1 heard sentences during training and test, listeners in Experiment 2 heard single words. Without access to the supra-segmental information present in sentences, listeners in Experiment 2 faced a more perceptually difficult task that required greater levels of attention to specific acoustic-phonetic details. Of the three training conditions, only transcription + talker ID yielded reliably more accurate test transcription performance than in the control condition (transcription of native-accented words during exposure). That only one of the training conditions (versus the three in Experiment 1) yielded learning confirms the relatively higher level of perceptual difficulty associated with understanding isolated accented words versus sentences. Transcription + talker ID is the only training condition in which listeners switched between transcription and talker-identification tasks and thus actively attended to both lexical and talker-specific details. The higher levels of generalized learning observed in this condition point to the importance of attention to both types of information in optimizing lexical processing to understand unfamiliar accented talkers and utterances.

Notably, test performance in the transcription condition in the current experiment was reliably lower than in the transcription + talker ID condition, despite comparable levels of task performance achieved by the end of the training phase for both easy and hard words (Fig. 3a and b). This was unexpected, as transcription of non-native-accented utterances during exposure has yielded robust perceptual learning at test in previous work (e.g., Bradlow & Bent, 2008; Tzeng et al., 2016) and in Experiment 1, though these paradigms employed sentence rather than word transcription tasks. One possible explanation for the seemingly detrimental effect of transcription observed in the current experiment is that single-word transcription hyper-focused listeners' attention on the acoustic-phonetic characteristics of the specific lexical items heard. Although this listening strategy yielded improved levels of comprehension of the voices and utterances presented during training, it precluded learning that generalized to *novel* voices and utterances at least with the amount of training or exposure provided in the current task. A second possibility is that listeners did not receive an adequate amount of exposure to the target accent to generalize their learning to novel utterances and talkers. Given that listeners in the current experiment heard word rather than sentence utterances and thus heard relatively fewer examples of accented utterances than in Experiment 1, it is possible that generalized learning would have occurred in Experiment 2 had the exposure phase been extended with additional tokens. With additional exposure trials, listeners might have been able to more effectively learn accent-general acoustic-phonetic characteristics of these single-word utterances.

Levels of learning at test also differed reliably between hard and easy words. For easy words, listeners achieved generalized learning only in the transcription + talker ID condition. For hard words, none of the three training conditions yielded reliably higher test performance relative to the control condition. In fact, test performance in the transcription condition was significantly lower than in control and transcription + talker ID conditions. Given evidence that phonological neighborhood density is negatively correlated with word recognition (Gahl & Strand, 2016), especially for non-native-accented speech (Chan & Vitevitch, 2015; Sidaras et al., 2009), the difficulty of transcribing words with many phonological competitors may have prevented listeners from extracting the necessary cues to generalize learning to previously unheard words spoken by unfamiliar voices. Nevertheless, results across both Experiment 1 and Experiment 2 in the current study collectively point to the significance of both lexical and indexical cues in understanding previously unheard non-native-accented voices.

General discussion

A wealth of empirical findings provide evidence for perceptual learning when listeners encounter non-native-accented speech. Still relatively unexplored are the cognitive mechanisms that allow listeners to understand voices that they have not heard before. The objective of the current work was to assess the extent to which attentional modulation during exposure to non-native-accented speech affects how perceptual learning generalizes to previously unheard talkers and utterances. Taken together, the results of the present study support the hypothesis that listeners' ability to adapt to accent-relevant properties of speech varies with shifts in attentional focus to relevant and distinct sources of variation in spoken language.

The novel contribution of this work lies in two major findings. First, orienting a listener toward indexical cues in non-native-accented speech offers a perceptual benefit for understanding previously unheard talkers and lexical items. Robust and comparable perceptual learning was observed in all exposure conditions (transcription, talker ID, transcription + talker ID) in Experiment 1, with exposure tasks that oriented listeners toward indexical cues yielding similar levels of generalized learning as those that oriented listeners toward linguistic properties. Second, when listeners had reduced access to global prosodic cues during exposure and learning was less constrained by the semantic and syntactic context of sentence-length utterances (Experiment 2), training that oriented listeners toward both indexical and lexical cues produced higher transcription accuracy than exposure that required listeners to attend only to one of the two types of information.

These findings are consistent with previous work showing that shifts in attentional focus change the extent to which perceptual learning generalizes to novel voices and utterances (Borrie et al., 2013; Loebach et al., 2008; Wright et al., 2015). That test performance varied at all as a function of condition in the current study suggests that directing listeners' attention to different channels of information in the speech signal changes which properties listeners encode during spoken language processing. Findings from both Experiment 1 and Experiment 2 point to the significance of attention to indexical cues in restructuring listeners' representations of variation in ways that promote the ability to understand unfamiliar voices with similar non-native-accented characteristics. Previous work has shown that familiarity with specific talkers' voices can improve talker intelligibility (e.g., Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000). The current findings extend this work by demonstrating that explicit attention directed toward a talker's identity can also facilitate the comprehension of previously unheard talkers from the same type of talker group (e.g., non-native-accented speech). Listeners in the transcription + talker ID condition in both experiments were provided with the opportunity to explicitly attend to both the lexical and indexical content of the presented stimuli. This attentional switch between two different types of informative cues perhaps optimized perceptual learning by facilitating the process of disambiguating talker-specific variation from accent-general systematicities.

It is worth noting that not all tasks that direct listeners to talker-related variation yield comparable levels of perceptual learning. Relative to asking listeners to complete a talker-identification task, asking listeners to identify the gender of the talker (Loebach et al., 2008), for example, resulted in less robust learning in a transcription task at test. Findings from the current study suggest that attention directed toward indexical cues that are unique to a specific talker's voice are especially useful, as they facilitate the disambiguation of talker-specific variation and variation that is characteristic of a speaker group (e.g., Spanish-accented talkers). The talker-identification task employed in exposure phase of the current study included talkers that differed in gender and accentedness. Future work may also systematically manipulate the acoustic similarity of the voices included in the exposure phase, as having to distinguish between acoustically similar voices may further sharpen listeners' ability to extract systematic accent characteristics.

Importantly, as talker identification is a markedly different task than the transcription task at test, the facilitative effect of talker-identification training in Experiment 1 cannot be attributed to task familiarity. For listeners in the talker ID condition, each exposure trial entailed that they identify the target talker from four possible response choices. Performance across the four sentence repetitions during the exposure phase suggests that listeners improved in their ability to distinguish among the four exposure talkers' voices. This improvement also implies an increased familiarity with the task demands. However, this task familiarity cannot account for the facilitative effect of talker-identification training on generalized learning, as listeners completed a transcription task at test. Unlike the talker-identification task completed during exposure, the transcription task required that listeners attend instead to the lexical characteristics of the spoken word stimuli. That listeners in the talker ID condition exhibited generalized perceptual learning to novel voices and utterances suggests that shifting listeners' attention to indexical characteristics allowed listeners to effectively register accent-based variation.

That generalized learning occurred only for the transcription + talker ID condition in Experiment 2 aligns with the results of previous work demonstrating the facilitative effect of shifting the listener's focus between indexical and lexical cues during exposure. In Sidaras et al. (2009) participants heard word-length utterances spoken by non-native English speakers during an exposure phase, switching between completing an accentedness rating task and a transcription task across blocks. Similar to the transcription + talker ID tasks employed in Experiment 2, listeners attended to both the idiosyncratic speech characteristics of each individual talker and the lexical characteristics that enable spoken word recognition. As in Sidaras et al. (2009) listeners in the current study who were provided with the opportunity to shift their attention between two sources of relevant cues exhibited robust generalized learning to previously unheard non-native-accented voices and words.

That levels of perceptual learning were highest in the transcription + talker ID condition in both experiments may be surprising given that switching between two different tasks could render the training experience more difficult for listeners. The potentially greater cognitive and perceptual load associated with switching between exposure tasks may provide a level of *desirable difficulty* (Bjork, 1994) that encouraged heightened listening effort (Van Engen & Peelle, 2014) and more elaborate processing. However, in general across the two experiments, perceptual learning as indexed by the generalization test did not appear to result from varying levels of desirable difficulty during training. Difficulty during training as indexed by initial performance on each exposure task, albeit with different performance measures, did not predict performance in the generalization test. Instead, relatively lower test performance in the transcription and talker ID conditions in Experiment 2 and relatively higher test performance with combined transcription and talker ID exposure did not seem to be attributable to the exposure phases presenting differentially difficult listening experiences (Gabay et al., 2017; Loebach et al., 2008).

The primary finding from Experiment 2 is that explicitly directing listeners' attention to *both* indexical and lexical details during exposure optimizes perceptual learning of non-native-accented speech. The facilitative effect of this exposure is greater than the effect of focusing listeners' attention on linguistic or indexical characteristics alone, underscoring the extent to which the processing of these two types of information is linked in speech perception (Nygaard & Pisoni, 1998; Perrachione et al., 2011; Remez et al., 1997). That is, the perceptual mechanisms that underlie talker identification are not independent from the mechanisms implicated in extracting the linguistic content of an utterance such that exposure and attention to properties associated with linguistic structure necessarily results in familiarity with properties associated with indexical variation. Linguistic and indexical information are thus not dichotomized. Rather, talker-specific information, especially when available from multiple talkers from the same speaker group (Sidaras et al., 2009; Tzeng et al., 2016), organizes listeners' linguistic speech sound representations in ways that encourage listeners to learn an accent independent of the vocal characteristics of an individual talker.

The current pattern of results aligns broadly with reward-based learning frameworks (e.g., Seitz et al., 2010; Vlahou et al., 2012; Wright et al., 2015). According to such models, perceptual learning occurs regardless of whether explicit attention is paid to the target features (accent characteristics shared among individuals in a speaker group), as long as the to-be-learned features are systematically paired with internal reward signals. These reward signals are elicited by successful performance on the target task and shift the perceptual system into a sensitized state that is especially conducive to learning. In the current study, the magnitude of the reward signal during each exposure trial varied across each training condition such that those who completed the talker ID or the transcription + talker ID tasks, relative to those who only completed the transcription task during exposure, were more likely to reach threshold levels of sensitization where sound categories could be modified to promote learning. This sensitized state may have been easier to reach in Experiment 1, as listeners heard sentence-length stimuli and thus had the opportunity to process informative variation on multiple dimensions, including acoustic-phonetic and supra-segmental details. In Experiment 2, when listeners heard single-word utterances during exposure, only listening conditions that provided sufficient informative variation to elicit above-threshold reward signals (transcription +talker ID) yielded generalized perceptual learning.

Distributional accounts of speech perception claim that listeners draw on the statistical contingencies between linguistic variability and talker- and group-specific factors to infer the speaker's intended message (Kleinschmidt & Jaeger, 2015; Kleinschmidt, 2019; McMurray & Jongman, 2016). According to such accounts, the listener tracks the distributions of phonemic categories within and across talkers and uses this experience to probabilistically infer what is being said. One virtue of these distributional accounts is that they account for both episodic and abstract representations of speech sounds such that as listeners accumulate instance-specific representations that include indexical details, distributions of these representations allow listeners to abstract away from individual episodes. Such models account for listeners' sensitivity to systematic variation at both the talker and group level. However, they do not address the cognitive mechanisms that enable listeners' registration of this variation. As the current results suggest, mapping spoken input to sound-category representations is attentionally guided (Heald & Nusbaum, 2014). A complete account of spoken language perception thus warrants an integration of computational and cognitive views.

Conclusion

Collectively, the current results show that attentional shifts between indexical and linguistic properties of speech modulate the extent to which listeners form talker-independent representations of non-native speech sounds. We suggest that attention paid to *both* linguistic and indexical cues optimizes the attribution of individual talker variation from group-based variation such that listeners can understand talkers and utterances that they have not encountered before. The current results constrain theoretical models of spoken language processing by (1) highlighting the role of attentional processes in modulating the way in which the speech stream is encoded and (2) underscoring the interdependency of indexical and linguistic cues in spoken language processing.

Open Practices Statement None of the data or materials for the experiments reported here is currently available, and none of the experiments was preregistered.

Data Availability Data and materials for the experiments reported can be made available upon request.

References

Alexander, J. E., & Nygaard, L. C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America, 145*(6), 3382–3398. https://doi.org/10.1121/1.5110302

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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. https://doi.org/10.1121/1.4789864

Bates D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing About Knowing* (pp. 185–205). MIT Press.

Borrie, S. A., McAuliffe, M. J., Liss, J. M., O'Beirne, G. A., & Anderson, T. J. (2013). The role of linguistic and indexical information in improved recognition of dysarthric speech. *The Journal of the Acoustical Society of America, 133*(1), 474–482. https://doi.org/10.1121/1.4770239

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Chan, K. Y., & Vitevitch, M. S. (2015). The influence of neighborhood density on the recognition of Spanish-accented words. *Journal of Experimental Psychology: Human Perception and Performance, 41*(1), 69–85. https://doi.org/10.1037/a0038347

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition, 108*(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *The Journal of the Acoustical Society of America, 102*(5), 2993–2996. https://doi.org/10.1121/1.420354

Drouin, J. R., & Theodore, R. M. (2022). Many tasks, same outcome: Role of training task on learning and maintenance of noise-vocoded speech. *The Journal of the Acoustical Society of America, 152*(2), 981–993. https://doi.org/10.1121/10.0013507

Flege, J., Schirru, C., & MacKay, I. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40*, 467–491. https://doi.org/10.1016/S0167-6393(02)00128-0

Gabay, Y., Karni, A., & Banai, K. (2017). The perceptual learning of time-compressed speech: A comparison of training protocols with different levels of difficulty. *PloS one, 12*(5), e0176488. https://doi.org/10.1371/journal.pone.0176488

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language, 89*, 162–178. https://doi.org/10.1016/j.jml.2015.12.006

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory & Cognition, 14*(3), 421–433. https://doi.org/10.1037/0278-7393.14.3.421

Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 35. https://doi.org/10.3389/fnsys.2014.00035

Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. *The Journal of the Acoustical Society of America, 131*(3), EL236–EL242. https://doi.org/10.1121/1.3685511

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104. https://doi.org/10.1121/1.397821

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means* (version R package version 1.2.4). https://CRAN.R-project.org/package=emmeans

Loebach, J. L., Bent, T., & Pisoni, D. B. (2008). Multiple routes to the perceptual learning of speech. *The Journal of the Acoustical Society of America, 124*(1), 552–561. https://doi.org/10.1121/1.2931948

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

McMurray, B., & Jongman, A. (2016). What comes after/f/? Prediction in speech derives from data-explanatory processes. *Psychological Science, 27*(1), 43–52. https://doi.org/10.1177/0956797615609578

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America, 85*(1), 365–378. https://doi.org/10.1121/1.397688

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon. *Research on Spoken Language Processing Report, 10*(3), 357–376.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*(3), 355–376. https://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 309–328. https://doi.org/10.1037/0278-7393.19.2.309

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Psychological Science, 333*, 595. https://doi.org/10.1126/science.1207327

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 651–666. https://doi.org/10.1037/0096-1523.23.3.651

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust, 17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Incorporated.

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition, 115*(3), 435–443. https://doi.org/10.1016/j.cognition.2010.03.004

Sidaras, S. K., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America, 125*(5), 3306–3316. https://doi.org/10.1121/1.3101452

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*, 1015. https://doi.org/10.3389/fpsyg.2013.01015

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25–B34. https://doi.org/10.1016/j.cognition.2005.01.006

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, and Psychophysics, 74*(6), 1284–1301. https://doi.org/10.3758/s13414-012-0306-z

Tzeng, C. Y., Alexander, J. E., Sidaras, S. K., & Nygaard, L. C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance, 42*(11), 1793–1805. https://doi.org/10.1037/xhp0000260

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*, 577. https://doi.org/10.3389/fnhum.2014.00577

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General, 141*(2), 363–381. https://doi.org/10.1037/a0025014

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, and Psychophysics, 75*(3), 537–556. https://doi.org/10.3758/s13414-012-0404-y

Wright, B. A., Baese-Berk, M. M., Marrone, N., & Bradlow, A. R. (2015). Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. *The Journal of the Acoustical Society of America, 138*(2), 928–937. https://doi.org/10.1121/1.4927411

Xie, X., & Myers, E. B. (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. https://doi.org/10.1016/j.jml.2017.07.005

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America, 143*(4), 2013–2031. https://doi.org/10.1121/1.5027410

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. https://doi.org/10.1037/xge0001039

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging, 15*(1), 88–99. https://doi.org/10.1037//0882-7974.15.1.88

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.