Published in final edited form as:

*J Exp Psychol Hum Percept Perform*. 2016 November ; 42(11): 1793–1805. doi:10.1037/xhp0000260.

# The Role of Training Structure in Perceptual Learning of Accented Speech

**Christina Y. Tzeng**,
Department of Psychology, Emory University

**Jessica E.D. Alexander**,
Department of Psychology, Centenary College of Louisiana

**Sabrina K. Sidaras**, and
Department of Psychology, Georgia State University

**Lynne C. Nygaard**
Department of Psychology, Emory University

## Abstract

Foreign-accented speech contains multiple sources of variation that listeners learn to accommodate. Extending previous findings showing that exposure to high-variation training facilitates perceptual learning of accented speech, the current study examines to what extent the *structure* of training materials affects learning. During training, native adult speakers of American English transcribed sentences spoken in English by native Spanish-speaking adults. In Experiment 1, training stimuli were blocked by speaker, sentence, or randomized with respect to speaker and sentence (Variable training). At test, listeners transcribed novel English sentences produced by Spanish-accented speakers. Listeners' transcription accuracy was highest in the Variable condition, suggesting that varying both speaker identity and sentence across training trials enabled listeners to generalize their learning to novel speakers and linguistic content. Experiment 2 assessed the extent to which ordering of training tokens by a single factor, speaker intelligibility, would facilitate speaker-independent accent learning, finding that listeners' test performance did not reliably differ across conditions. Overall, these results suggest that the structure of training exposure, specifically trial-by-trial variation on both speaker's voice and linguistic content, facilitates learning of the systematic properties of accented speech. The current findings suggest a crucial role of training structure in optimizing perceptual learning. Beyond characterizing the types of variation listeners encode in their representations of spoken utterances, theories of spoken language processing should incorporate the role of training structure in learning lawful variation in speech.

### Keywords

perceptual learning; speech perception; foreign-accented speech; training structure

Correspondence concerning this article should be addressed to Christina Y. Tzeng, 36 Eagle Row, Department of Psychology, Emory University, Atlanta, GA 30322; ctzeng@emory.edu.

Speech is a highly variable signal that changes both within and across speakers. Realization of a linguistic utterance varies as a function of numerous factors, including differences in phonetic context (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), individual patterns of articulation (Ladefoged, 1989), and region of speaker origin (Peterson & Barney, 1952). Despite this variation, listeners perceive stable linguistic units. Much empirical and theoretical work has been done to characterize the *types* of variation that are encoded in listeners' representations of speech. The primary objective of the present research is to clarify the *processes* by which listeners overcome this variation. In particular, we assess how the *structure* of previous experience with highly variable speech can facilitate spoken language comprehension.

Traditional approaches have assumed that during speech processing, variation due to surface characteristics is eliminated or normalized (e.g., Ladefoged & Broadbent, 1957; Miller, 1989). For perceptual accommodation to occur, acoustic variation, such as that due to differences in speakers' voices, is discarded. In these speaker-normalization approaches, the speech signal is recalibrated or normed to align with canonical representations stored in memory (Cutler, 2008; Cutler, Eisner, McQueen, & Norris, 2010). However, a growing number of studies have provided converging evidence that surface characteristics of spoken utterances, such as speaker's voice, speaking rate, and intonation contour, are encoded in listeners' memory representations of spoken utterances (Goldinger, 1996; 1998; Palmeri, Goldinger, & Pisoni, 1993). During spoken word recognition, listeners may map each spoken word to episodic traces of previously encountered spoken words such that variation in linguistic structure is preserved during the representation and processing of spoken language (Goldinger, 1996; 1998).

Evidence that speaker-dependent variation is encoded during spoken word recognition suggests that one way listeners may achieve stable linguistic percepts is through a perceptual adaptation process by which listeners update their representations to align with particular types of variation in the speech input (McQueen, Cutler, Norris, 2006; Sjerps & McQueen, 2010). Several studies have shown that familiarity with speaker variation facilitates speech processing (e.g., Allen & Miller, 2004; Bradlow & Pisoni, 1999; Eisner & McQueen, 2005; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Yonan & Sommers, 2000) and may restructure linguistic representation (e.g., Creel, Aslin, & Tanenhaus, 2008; Dahan, Drucker, & Scarborough, 2008; Kraljic, Samuel, & Brennan, 2008; Norris, McQueen, & Cutler, 2003; Trude & Brown-Schmidt, 2012). Nygaard et al. (1994), for example, found that after learning to identify different speakers during training, listeners performed more accurately on a word recognition test if the words were spoken by familiar rather than unfamiliar speakers, suggesting that speaker-dependent variation is encoded by listeners and affects word recognition.

Foreign-accented speech provides unique opportunities to investigate the processes by which listeners overcome variation in the speech signal, as accented speech differs from listeners' native pronunciations along multiple acoustic and phonetic dimensions. Unfamiliar sound-to-meaning mappings in accented speech can result in lexical ambiguity for native listeners and often require increased processing effort (Munro, 1993). However, because non-native speakers' accented productions reflect the phonology and articulatory patterns of their native

language, the acoustic and phonetic features of accented speech are systematic across speakers from the same language background, or *accent-general* (e.g., Flege, Bohn, & Jang, 1997; Sidaras, Alexander, & Nygaard, 2009). Critical for comprehension of accented speech is the ability to distinguish between speaker-dependent and speaker-independent variation, which co-occur in the speech signal.

Prior research suggests that listeners can learn to distinguish these two sources of variation to achieve speaker-independent adaptation to accented speech (e.g., Kraljic & Samuel, 2005, 2006; Lively, Logan, & Pisoni, 1993). In particular, this work has shown that high speaker variation facilitates cross-speaker generalization by allowing listeners to distinguish speaker-specific idiosyncrasies from speaker-independent regularities of accented speech. (Barcroft & Sommers, 2005; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Greenspan, Nusbaum, & Pisoni, 1988; Lively et al., 1993). Bradlow and Bent (2008), for example, examined to what extent exposure to multiple speakers promotes learning of accent-general properties of foreign-accented speech. During training, native English-speakers transcribed sentences spoken by either one or multiple accented speakers. At test, those who heard sentences spoken by multiple speakers transcribed sentences spoken by a novel accented speaker more accurately than those who only heard one speaker during training, suggesting that exposure to multiple speakers facilitates speaker-independent learning of accent. Bradlow and Bent's (2008) findings align with results from other studies demonstrating a benefit of variable training. For example, Lively et al. (1993) found that Japanese listeners learned the English /r/-/l/ contrast when presented with training stimuli spoken either by multiple speakers or by a single speaker. However, only listeners presented with highly variable training sets generalized their learning to tokens produced by a novel speaker.

Sidaras et al. (2009) assessed the extent to which highly variable training encourages perceptual adaptation to multiple novel speakers. Native English speakers transcribed words or sentences spoken by Spanish-accented speakers during training. Transcription performance of novel tokens produced by both familiar and unfamiliar speakers was reliably more accurate for listeners who completed the training phase than those who did not, suggesting that exposure to multiple speakers and lexical items encourages listeners to distinguish speaker-dependent from speaker-independent variation to learn systematic properties of a particular accent. Highly variable training thus allows listeners to identify ways in which the sound structure of foreign-accented speech systematically differs from that of native-accented speech. The ability to distinguish between relevant sources of variation, speaker-dependent and speaker-independent, during accent learning is critical for effective accommodation to unfamiliar accented speech tokens.

The literature discussed thus far provides evidence that listeners readily adapt to foreign accented speech. Further, these findings suggest that high-variation training environments facilitate the disambiguation of speaker-dependent from speaker-independent characteristics in accented speech. However, why this variation is beneficial remains unclear. Exemplar-based models assume that high-variation exposure increases the number of memory traces associated with a particular utterance and thus creates a particularly comprehensive representation of a given accent category (Goldinger, 1996; 1998). One possibility that builds on exemplar views is that exposure to multiple speakers and linguistic contexts allows

listeners to compare tokens to identify the similarities and differences across them and distinguish relevant from irrelevant variation, thus inducing category structure for accent-specific regularities (Gentner & Markman, 1994). The role of comparison has been examined in the categorization of novel objects (Gentner & Namy, 1999; Namy & Gentner, 2002), artistic styles (Kang & Pashler, 2012; Kornell & Bjork, 2008), and in face recognition (Clutterbuck & Johnston, 2005; Dwyer & Vladeanu, 2009). Kornell and Bjork (2008), for example, found that participants who saw paintings by different artists interleaved (spaced) during training better identified novel paintings by these artists than participants who saw the paintings blocked by artist (massed). This finding suggests that spacing of training tokens allowed for more effective discrimination of relevant category features.

Learning to disambiguate systemic accent characteristics from speaker-dependent variation may involve comparison processes that facilitate identification of relevant acoustic and phonetic features of an accent category. Speaker-independent accent learning occurs once the listener effectively identifies which features are consistent across speakers and which are idiosyncratic to particular individuals. One potentially important aspect of this category learning process, as proposed by Gentner (1983), is that comparison of highly similar, prototypical exemplars increases the salience of relevant cues for deciding category membership and facilitates categorization of more variable, less prototypical exemplars (Gentner, 1983; Gentner & Medina, 1998). Known as the *progressive alignment effect*, this "easy-to-hard" learning process may apply for accent learners such that exposure to accented speech tokens that are more easily understood may facilitate the categorization of tokens that are relatively less prototypical or intelligible. Thus, the order in which category exemplars are presented during training may affect the robustness of category learning.

That listeners are more likely to generalize learning to novel speakers after highly variable training implies that accent leaning may involve comparison and progressive alignment of accented tokens. However, few studies have examined this possibility. Sumner (2011) found that native English speakers who heard French-accented /b/-initial and /p/-initial words with variable VOTs experienced greater boundary shifts in their subsequent categorization of /ba/-/pa/ syllables relative to listeners who heard invariant VOTs during the exposure phase, suggesting that variable exposure may lead to shifts in category structure to incorporate non-native pronunciations. Within the variable VOT exposure conditions, the order of word tokens was manipulated such that listeners either heard tokens presented randomly, progressively from more native-like to less native-like, or less native-like to more native-like. Categorization performance among these three conditions varied such that listeners who heard more native-like tokens first experienced the largest boundary shifts, suggesting that alignment of accented tokens on a particular dimension or series of dimensions may encourage a comparison process that facilitates understanding of unfamiliar pronunciations.

Taken together, the described findings suggest that for perceptual learning of accent, the structure of the training exposure may affect listeners' ability to identify speaker-dependent versus speaker-independent characteristics of accented speech. The purpose of the present study was to assess this possibility. Exemplar-based models of perceptual learning (e.g., Goldinger, 1996; 1998) currently do not account for the role of training structure, as these views suggest only that the amount rather than the organization of previous experience

affects the robustness of learning. Thus, findings suggesting that the structure of training may indeed affect accented speech comprehension would potentially qualify the extent to which exemplar-based theories can account for the nature of the perceptual learning process. In the current study, we consider the possibility that theories predicting the facilitative effects of spaced versus massed training (e.g., Cepeda et al., 2006; Fine & Jaeger, 2013; Glenberg, 1979) may apply to perceptual learning of accented speech.

In the present study, we aimed to equate the *amount* of variation in the training stimuli presented across conditions, with conditions differing only in the *organization* of the training materials such that listeners heard tokens blocked by speaker, sentence, or randomized with respect to both speaker and sentence. In two experiments, English-speaking listeners completed training sessions in which they heard Spanish-accented speech produced by multiple non-native speakers. At test, listeners transcribed utterances produced by unfamiliar accented speakers. The objective of these studies was to examine the extent to which the organization of training materials would encourage comparison and alignment processes that would promote generalizable accent learning.

If training structure affects the robustness of generalizable accent learning, then test performance may vary as a function of the presentation order of tokens during training. If accent learning depends simply on exposure to a given range of type/token pairings during training, then the grouping or ordering of items during training should have no influence on the robustness of accent learning as indexed by generalization to novel talkers. Although exemplar-based models would predict that exposure to systematic variation would induce learning of a non-native accent, these models currently do not account for potential differences in learning due to training structure or ordering of exemplars. If training structure does influence accent learning, then it would suggest that listeners are not only encoding variation but also engaging in a perceptual process that may involve comparison of similarities and differences across tokens to identify speaker-independent from speaker-dependent sources of variation.

Manipulating the nature of trial-to-trial changes in speaker and linguistic content across training conditions will also allow for an examination of the type of processes that may underlie any effect of differences in training structure. For example, for conditions in which items are blocked by speaker, there is trial-to-trial variation in linguistic content but not in speaker's voice. If this type of variation during training encourages the most robust learning as indexed at test, then it would suggest that the opportunity to attend to and compare how each speaker produces a range of linguistic structures may facilitate identification of speaker variation and encourage separation of speaker from accent properties. For conditions in which items are blocked by linguistic content, there is trial-to-trial variation in speaker's voice but not in linguistic content. Robust learning under these conditions would suggest that the opportunity to attend to and compare how particular linguistic structures vary across speakers may allow for better identification of the properties of accented speech. However, if exposure to trial-to-trial changes in *both* dimensions results in highest test performance, then it would suggest that listeners may need exposure to variation along both speaker identity and linguistic content during perceptual learning of accent.

To examine the potential role of progressive alignment in learning accent properties, the order of speaker blocks was manipulated during training (Experiment 2). Training tokens were blocked by speaker, but speaker order was manipulated such that listeners either encountered training trials first from high or low intelligibility speakers, or from interleaved high and low intelligibility speakers. If initial training with high- rather than low-intelligibility speakers results in highest transcription accuracy at test, it would suggest that alignment of accented tokens by speaker intelligibility facilitates the extraction of commonalities across speakers and encourages speaker-independent perceptual learning of accent.

## Experiment 1

Experiment 1 assessed the extent to which organization of accented training tokens would affect perceptual learning of foreign accented speech. During training, listeners transcribed sentences spoken by Spanish-accented speakers. Across training conditions, organization of tokens varied such that listeners heard sentences blocked by speaker, sentence, or randomized with respect to both speaker and sentence. Critically, listeners heard the same set of tokens in each training condition, with only the order in which the tokens were presented varying across conditions. At test, listeners transcribed novel sentences spoken by unfamiliar accented speakers.

Assessing the effects of organization of training tokens will clarify how trial-by-trial variation results in optimal perceptual learning. Given previous findings that robust and generalizable perceptual learning requires exposure to multiple speakers and phonetic contexts (Lively et al., 1993; Sidaras et al., 2009), we predicted that trial-by-trial variation in both speaker identity and sentence item (Variable training) would encourage the most robust learning. Variation across both dimensions would allow listeners to readily distinguish speaker-dependent idiosyncratic pronunciations from speaker-independent systematicity in accented speech.

### Method

**Participants—**One-hundred and forty-five Emory University undergraduates either received course credit (n = 74) or were paid $15 for their participation (n = 71). All were native English-speaking monolinguals[1] and reported no history of speech or hearing disorders.

**Materials—**Stimuli used in this experiment were a subset of those used in Sidaras et al. (2009). Eight native Spanish speakers (four male, four female) from Mexico City were recruited from the Atlanta area to record word and sentence stimuli. Each speaker produced 144 monosyllabic English words and 100 Harvard sentences (Rothauser et al., 1969). Sentences ranged from 6 to 10 words and were phonetically balanced to reflect the frequency of phonemes in English. Recordings were re-digitized at a 22050 Hz sampling

[1]Only individuals who (1) had learned English as their first language and (2) were not fluent in any other language were eligible to participate in the current experiments. Native English-speaking monolingual status for participants was verified using responses from a language background questionnaire that participants completed prior to the experimental task.

rate, edited into separate files, and amplitude normalized. Only the sentences were used in the current study.

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 eight speakers (10 listeners per accented speaker). The proportion of correctly transcribed words was calculated across listeners for each of the eight speakers ($M$ = 81.59%). An additional 10 listeners rated the accentedness of ten sentence-length utterances from each of the eight speakers. Listeners rated the accentedness of each sentence on a 7-point Likert-type scale, from 1 = "not accented" to 7 = “very accented”. Mean accentedness ratings were calculated for each speaker ($M$ = 4.38, range = 3.10 – 6.17). Table 1 lists mean accentedness ratings, as well as baseline sentence intelligibility scores for each speaker.

Stimuli used in the current study consisted of 56 of the Harvard sentences. Two groups of four speakers (two male and two female) were constructed to create training and test groups. Speaker groups were equated overall for sentence intelligibility ($M_1$ = 80.45, $M_2$ = 82.73) and accentedness ($M_1$ = 4.07; $M_2$ = 4.70) 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 speakers (one male and one female) were characterized as high intelligibility ($M_1$ = 90.15; $M_2$= 87.25), and two as low intelligibility speakers ($M_1$ = 70.75; $M_2$ = 78.20). Across both speaker groups, intelligibility ratings were reliably higher for high-intelligibility versus low-intelligibility speakers, $t(6)$ = 4.12, $p$ = .006, $M_{high}$ = 88.70; $M_{low}$ = 74.48.

**Procedure—**Listeners were assigned to one of four conditions, three of which included both training and test phases. Speaker group was counterbalanced for all training conditions such that half the listeners in each condition heard Group 1 during training, and half heard Group 2 during training. Listeners completed the experiment in groups of four or fewer, with stimulus presentation controlled on a PC computer using E-prime 1.1 (Schneider, Eschman, & Zuccolotto, 2002). Auditory stimuli were presented binaurally over Beyerdynamic DT100 headphones at approximately 75 dB SPL.

**Training phase:** During training, listeners heard 36 sentences spoken four times each, once by each of four Spanish-accented speakers. Upon hearing each sentence, listeners transcribed what they heard. After each response, listeners saw the target sentence presented on the computer screen and then heard the sentence a second time. In other words, the training phase was supervised such that listeners received explicit feedback followed by stimulus repetition.

Across the three training conditions, listeners transcribed the same tokens. However, the presentation order of tokens varied across training conditions, allowing us to assess the effect of training structure on accuracy of learning. Listeners in the Sentence condition heard tokens grouped by sentence such that they heard each sentence repeated four times in succession, once by each speaker, with speaker and sentence order randomized. Listeners in the Speaker condition heard sentences grouped by speaker, such that they heard all 36 sentences spoken by one speaker before hearing the same sentences spoken by the next

speaker, with speaker order systematically manipulated such that listeners heard high-intelligibility speakers first (HHLL), last (LLHH), or interleaved (LHLH, HLHL). Listeners in the Variable condition heard the same set of sentences and sentence-speaker pairings, but items were randomized with respect to both sentence and speaker such that speaker's voice and sentence content both varied from trial-to-trial. Those assigned to the Control condition received no training and completed only the test phase.

**Generalization Test:** To examine the extent to which perceptual learning generalized to novel speakers and sentential content, listeners in all conditions heard 20 novel sentences at test spoken by four Spanish-accented speakers that they did not hear during training. Test sentences were mixed in white noise (+10 signal-to-noise ratio), with sentence-speaker pairings randomized. Listeners transcribed five sentences spoken by each of the four speakers without corrective feedback.

**Analysis—**Analyses were conducted in R (version 3.0.3; R Core Team, 2014) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Follow-up comparisons were conducted using the multcomp package (Hothorn, Bretz, & Westfall, 2008). Logistic mixed-effects models were chosen over traditional means comparisons[2] (e.g., ANOVA) to (1) represent all systematic sources of variance in our outcome variables by accounting for both fixed and random effects and (2) maximize the statistical power of our analyses (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tilly, 2013; Jaeger, 2008; Locker, Hoffman, & Bovaird, 2007; Raaijmakers, 2003).

## Results and Discussion

**Training phase—**Across conditions, participants heard and transcribed each of the 36 training sentences four times, once by each speaker. The proportion of correctly transcribed words in each sentence was calculated for all sentences. Each word was coded as either correct (1) or incorrect (0). Homophones (e.g., creak, creek), regular verb tense changes (e.g., cook, cooked) and clearly identifiable words that contained minor typographical errors (e.g., chicken, chiken) were scored as correct. Words that were misspelled and ambiguous (e.g., thin, tind) were coded as incorrect.

Fig. 1 shows the transcription accuracy for each of the four times the sentences were repeated. A logistic mixed-effects model assessed the extent to which transcription accuracy varied as a function of repetition and condition. Sentence and subject were included as random effects with random slopes allowing subjects to vary with respect to repetition. Best-fitting models for all analyses were determined using step-wise model comparisons using log-likelihood ratio tests (Baayen et al., 2008). Including the interaction between repetition and condition in the model significantly improved model fit, $\chi^2$ (6) = 17.20, p = 0.008, suggesting that the trajectory of learning across sentence repetitions differed as a function of condition. Planned contrast analyses indicated that transcription accuracy in the Variable,

[2]In addition to logistic mixed-effects models, linear mixed-effects models and ANOVAs were also run for all reported analyses across all experiments. Results patterned similarly across all three statistical approaches. Cases where statistical results for test performance deviated across approaches are noted. Results from logistic rather than linear mixed models are reported as the former more accurately accounts for variance in data that are bounded by 0 and 1 (Barr, 2008; Jaeger, 2008).

Sentence, and Speaker conditions differed as a function of repetition. Reliable improvement in transcription accuracy was observed between repetitions 1 and 2 (Variable, $b = 1.88$, $p < .001$; Sentence, $b = 2.41$, $p < .001$; Speaker, $b = 1.13$, $p < .001$). For the Variable condition, transcription accuracy also improved significantly in the Variable condition between repetitions 3 and 4, $b = 2.09$, $p < .049$, but decreased significantly in the Sentence condition, $b = -2.21$, $p = .037$. Although participants' learning trajectories differed as a function of condition, contrast analyses indicated that listeners' transcription accuracy across conditions was significantly higher for repetition 4 ($M = .98$, $SD = .04$) than for repetition 1 ($M = .90$, $SD = .04$; $b = 2.33$, $p < .001$), suggesting that participants engaged in perceptual learning for all conditions in the training phase.

**Generalization test—**It was hypothesized that training would result in increased transcription accuracy and that training that highlighted speaker-independent properties of accent (Variable training) would result in more accurate transcription than training that highlighted speaker-specific properties (Speaker training) or sentence-specific properties (Sentence training). Fig. 2 shows transcription accuracy (proportion of correctly transcribed words) as a function of condition. A logistic mixed-effects model assessed the extent to which transcription accuracy varied as a function of condition. Sentence and subject were included as random effects. Including condition in the model as a fixed effect significantly improved model fit, $\chi^2$ (3) = 12.18, p = 0.007, suggesting that test performance differed reliably across conditions. Contrast analyses indicated significantly higher transcription accuracy in the Variable condition than in the Control condition, $b = 0.29$, $p = .046$, Sentence condition, $b = 0.33$, $p = .020$, and Speaker condition, $b = 0.48$, $p < .001$. All other comparisons were not significant (all $b$s < .19, all $p$s > .156).

Transcription accuracy at test was highest for listeners who completed Variable training, suggesting that trial-by-trial changes in both speaker identity and linguistic content encouraged generalized learning to novel speakers and sentences. Given that Variable training included the most trial-to-trial change and may be perceived as the most difficult of the training conditions, optimal test performance in this condition may be surprising (but see Alter, Oppenheimer, Epley, & Eyre, 2007; Bjork, 1994). However, trial-by-trial change across two variables may have allowed listeners in the Variable condition to more readily separate variation due to acoustic differences in speakers' voice from systematic variation due to accent. Grouping tokens by sentence or speaker highlighted *either* pronunciations with limited exposure to varying phonetic environment *or* idiosyncratic characteristics of speakers' voices. Neither of these two training structures provided optimal conditions under which listeners could extract speaker-independent systematicities of accent.

Taken together with the finding that test performance in the Sentence and Speaker conditions did not reliably differ from that in the Control condition, the facilitative effect of Variable training suggests that optimal accent learning occurs when both linguistic content and speaker identity change together during training. However, one alternative possibility is that this facilitative effect is instead driven by the similarity in task strategy employed by listeners at Variable training and at test. In both, listeners transcribed sentences that varied by speaker and sentence across successive trials. Thus, rather than learning speaker-independent accent regularities, listeners may have instead learned a task-specific attentional

strategy for coping with trial-to-trial changes in speaker and sentence. Although some (e.g., Sidaras et al., 2009) have found no difference in test performance between English training, where trial to trial changes occur with accented exposure, and no training control conditions, others (e.g., Bradlow & Bent, 2008) have found that test accuracy is higher in English training versus no training conditions, suggesting some benefit for task familiarity. The facilitative effect of Variable training in our test may thus have occurred as a consequence of adaptation to task strategy rather than a consequence of exposure to informative variation across trials. Experiment 1b addresses this possibility.

## Experiment 1b

The objective of Experiment 1b was to assess the extent to which reliably better test performance in the Variable training condition could be attributed to the similarity between task strategy employed during training and test. In the current experiment, we compared listeners' transcription accuracy at test after completing Variable training identical to that implemented in Experiment 1 (Spanish Variable), Variable training with sentences spoken by *native* English speakers (English Variable), or no training (Control). If the facilitative effect of Variable training occurs as a result of task familiarity, test performance should not differ between the two training conditions. If the facilitative effect is instead due to experience with informative variation that encourages generalized learning to novel accented speakers and items, test performance in the Spanish Variable condition should be reliably better than in the English Variable and no training conditions. For *both* Spanish Variable and English Variable training, listeners must attend to changes in speaker identity and sentence from trial to trial. Thus, higher test performance in the Spanish Variable condition relative to that in the English Variable condition would suggest that the facilitative effect of Variable training cannot be entirely attributed to listeners' implementation of a task-specific attentional strategy.

### Method

**Participants—**Sixty Emory University undergraduates received course credit for their participation. All were native English-speaking monolinguals and reported no history of speech or hearing disorders.

**Materials—**Stimuli presented during training in the Spanish Variable and at test in all conditions were identical to those used in Experiment 1. Training stimuli in the English Variable condition consisted of the same 56 Harvard Sentences as in the Spanish Variable condition but spoken instead by eight (four male and four female) native English speakers who spoke a Standard American English accent familiar to the participants.

**Procedure—**Listeners were assigned to the Spanish Variable, English Variable, or Control condition. Training and test structure were identical in the Spanish Variable and English Variable conditions such that listeners transcribed 36 sentences spoken four times each, once by each of the four speakers, with speaker and sentence varying with each trial. At test, listeners in all conditions transcribed 20 novel sentences mixed in white noise spoken by

Spanish-accented speakers who were not heard during training. All other aspects of the procedure were identical to that in Experiment 1.

## Results and Discussion

**Training phase—**The proportion of correctly transcribed words in each sentence was calculated for all sentences. Fig. 3 shows the transcription accuracy for each of the four times the sentences were repeated in both the Spanish and English Variable training conditions. A logistic mixed-effects model assessed the extent to which transcription accuracy varied as a function of repetition and condition. Sentence and subject were included as random effects with random slopes allowing subjects to vary with respect to repetition. Including the interaction between repetition and condition in the model significantly improved model fit, $\chi^2$ (3) = 14.73, p = 0.003, suggesting that the trajectory of learning across sentence repetitions differed as a function of condition. Planned contrast analyses indicated that transcription accuracy in the English Variable condition did not differ as a function of repetition. Performance in the Spanish Variable condition differed significantly across repetitions with reliable improvement between repetitions 1 and 2, $b$ = 1.14, $p$ < .001, and between repetitions 3 and 4, $b$ = 2.15, $p$ < .001. For Spanish Variable training, listeners' transcription accuracy was significantly lower for the first ($M$ = .88, $SD$ = .03) versus last ($M$ = .98, $SD$ = .03) repetition, $b$ = −2.84, $p$ < .001, suggesting that perceptual learning occurred. For English Variable training, accuracy for the first ($M$ = .97, $SD$ = .02) versus last ($M$ = .98, $SD$ = .02) did not differ, $b$ = −0.20, $p$ = .818, as listeners' accuracy was already near ceiling when transcribing the sentences at repetition 1.

**Generalization test—**Fig. 4 shows the proportion of correctly transcribed words at test as a function of condition. A logistic mixed-effects model assessed the extent to which transcription accuracy varied as a function of condition. Sentence and subject were included as random effects. Including condition in the model as a fixed effect significantly improved model fit, $\chi^2$ (2) = 20.12, $p$ < .001, suggesting that test performance differed reliably across conditions. Contrast analyses indicated significantly higher transcription accuracy in the Spanish Variable condition than in the English Variable condition, $b$ = 0.46, $p$ = .008, and the Control condition, $b$ = 0.78, $p$ < .001. The comparison between performance in the English Variable and Control conditions was also significant, with reliably higher transcription accuracy in the English Variable condition, $b$ = 0.32, $p$ = .034.[3]

This pattern of findings suggests that adaptation to the task, which involved learning trial-to-trial variation in speaker's voice and linguistic content for both Spanish and English training conditions, cannot fully account for superior test performance in the Variable condition in Experiment 1, as transcription accuracy at test for the Spanish Variable condition was reliably higher than for both the English Variable and Control conditions. Reliably higher performance in the English Variable condition relative to that in the Control condition suggests that familiarity with task strategy does offer some benefit at test. However, if the

[3]The contrast comparisons assessing differences between test performance in the English Variable and Control conditions yielded no significant difference between the two conditions in the *linear* mixed model, $b$ = .02, $p$ = .348, and the ANOVA follow-up comparisons, $p$ = .352,. The reliable difference found with the *logistic* mixed model at $p$ = .034 should thus be interpreted with caution.

facilitative effect of Variable training in Experiment 1 occurred *solely* as an artifact of learning task strategy, rather than speaker-independent accent regularities, then test performance in the English and Spanish Variable training conditions in the current experiment would not have differed, as both conditions required that listeners transcribe sentences that varied trial-to-trial by speaker identity and sentence.

A second possibility that may account for relatively higher performance in the Variable condition in Experiment 1 is that listeners in the Variable condition might have been able to more easily apply their speaker-independent accent knowledge at test because the structure of the test phase was similar to the structure of the training phase. However, this possibility is unlikely given related findings from previous work. Bradlow and Bent (2008) compared listeners' transcription performance of accented sentences spoken by a single novel speaker after hearing multiple accented speakers or a single accented speaker during training. Transcription performance at test was higher in the multiple- versus single-speaker condition, suggesting that similarity in task structure cannot fully account for generalization of learning, as the task structure in the multiple-speaker condition differed between training and test. Whereas training in the multiple speaker condition varied by both speaker and sentence from trial to trial, listeners only heard one speaker at test (blocked presentation). If perceptual learning of accent were driven primarily by similarity in task structure, listeners would not have exhibited learning in the multiple-speaker condition. In the current study, mismatching task structure between training and test is unlikely to fully account for the lack of evidence for learning in the Speaker and Sentence training conditions. Taken together, the findings from Experiments 1 and 1b suggest that speaker-independent learning of accent occurs when listeners have the opportunity to compare across relevant sources of variation from trial-to-trial.

## Experiment 2

Having established in Experiment 1 that training structure affects learning and generalization, Experiment 2 explored the unexpected finding that no evidence of learning was found in the Speaker condition in Experiment 1. Given this unexpected result, one hypothesis is that the ordering of speakers during training affects accent learning. Previous evidence suggests that systematic ordering of training materials by a particular dimension may facilitate listeners' understanding of unfamiliar non-native pronunciations (e.g., Sumner, 2011). Thus, Experiment 2 examined the effects of Speaker training in the context of a particular type of training structure, progressive alignment. Specifically, Experiment 2 assessed the extent to which progressive alignment of training tokens by a single variable, speaker intelligibility, would facilitate speaker-independent accent learning.

Speaker intelligibility has been found to be associated with acoustic-phonetic properties, such as vowel space dispersion (Bradlow, Torretta, & Pisoni, 1996; Bond & Moore, 1994) and consonant-to-vowel amplitude ratios (Hazan & Simpson, 1998), suggesting that intelligibility may index acoustic-phonetic and other properties (e.g., prosodic contours) that are more or less native-like. Bradlow and Bent (2008) found faster adaptation to relatively high versus low intelligibility foreign-accented speech, suggesting that the more consistent and accurate pronunciations found in high-intelligibility utterances might facilitate

perceptual learning of accent. Just as progressive alignment of tokens is generally hypothesized to facilitate a comparison process that may highlight commonalities and differences among exemplars (Gentner, 1983; Gentner & Medina, 1998), progressive alignment of accented tokens by speaker intelligibility may facilitate the extraction of commonalities across speakers and encourage speaker-independent perceptual learning of accent. In the current experiment, conditions that included progressively aligned stimuli involved presenting accented sentences spoken by highly intelligible followed by those that were progressively less intelligible, thus creating an "easy-to-hard" training structure.

Listeners in Experiment 2 completed the same training as in the Speaker condition from Experiment 1 with the order of high- and low-intelligibility speaker blocks systematically manipulated.[4] If ordering tokens by speaker intelligibility allows listeners to more readily identify speaker-independent acoustic properties of accent, then we would predict that listeners will achieve highest mean transcription accuracy at test when trained with high- rather than low-intelligibility speakers first.

## Method

**Participants—**One-hundred and twenty-three Emory University undergraduates either received course credit (n = 61) or were paid $15 for their participation (n = 62). All were native English speakers and reported no history of speech or hearing disorders.

**Materials—**Stimuli were identical to those used in Experiment 1.

**Procedure—**Listeners were randomly assigned to one of four conditions that included both training and test sessions. All other aspects of the procedure were identical to that in Experiment 1.

**Training phase:** During training, listeners heard 36 sentences spoken four times each, once by each of four Spanish-accented speakers. As in the Speaker condition in Experiment 1, listeners heard all 36 sentences spoken by one speaker before hearing the same sentences spoken by the next speaker. The order of high- and low- intelligibility speaker blocks was systematically manipulated such that listeners heard high-intelligibility speakers first (HHLL), last (LLHH), or interleaved (LHLH, HLHL). Systematic ordering of high- and low-intelligibility speakers allowed for examination of order effects along this speaker characteristic.

**Generalization test:** Listeners completed a test phase identical to that in Experiment 1.

## Results and Discussion

**Training phase—**The proportion of correctly transcribed words in each sentence was calculated for all sentences. Fig. 5 shows the transcription accuracy for each of the four

[4]Although the training conditions in Experiment 2 were identical to the Speaker training condition in Experiment 1, a separate group of participants were run in order to independently assess the extent to which progressive alignment by speaker intelligibility would affect perceptual learning. Rather than include additional participants in the Speaker condition of Experiment 1 and conducting a follow-up analysis, this approach ensured that the number of participants in Experiment 1 remained constant across the training conditions.

times the sentences were repeated. A logistic mixed-effects model assessed the extent to which transcription accuracy varied as a function of repetition and condition. Sentence and subject were included as random effects with random slopes allowing subjects to vary with respect to repetition. Best-fitting models for all analyses were determined using step-wise model comparisons using log-likelihood ratio tests. Including the interaction between repetition and condition in the model significantly improved model fit, $\chi^2$ (9) = 115.31, p < 0.001, suggesting that the trajectory of learning across sentence repetitions differed as a function of condition.

Planned contrast analyses indicated that transcription accuracy in the LLHH condition differed as a function of repetition with reliable improvement in performance between repetitions 1 and 2, $b$ = 1.38, $p$ < .001, and between repetitions 2 and 3, $b$ = 1.71, $p$ = .005. Performance in the HHLL condition reliably improved between repetitions 1 and 2, $b$ = 1.12, $p$ = .017, and a reliably decreased between repetitions 2 and 3, $b$ = −1.16, $p$ = .017. For the LHLH condition, transcription accuracy reliably improved between repetitions 1 and 2, $b$ = 0.10, $p$ < .001, and between repetitions 3 and 4, $b$ = 0.02, $p$ < .001, and reliably decreased between repetitions 2 and 3, $b$ = −0.01, $p$ = .025. Lastly, transcription accuracy in the HLHL condition reliably improved between repetitions 1 and 2, $b$ = 0.009, $p$ = .041, and between repetitions 2 and 3, $b$ = 0.04, $p$ < .001, and a reliably decreased between repetitions 3 and 4, $b$ = −0.01, $p$ = .022. Although participants' learning trajectories differed as a function of condition, contrast analyses indicated that listeners' transcription accuracy across all conditions was significantly higher for repetition 4 ($M$ = .98, $SD$ = .03) than for repetition 1 ($M$ = .91, $SD$ = .05; $b$ = 1.72, $p$ < .001), suggesting that participants engaged in perceptual learning across all conditions in the training phase.

**Generalization test—**Fig. 6 shows transcription accuracy (proportion of correctly transcribed words) as a function of condition. A logistic mixed-effects model assessed the extent to which transcription accuracy varied across the four training conditions and the Control condition from Experiment 1. Sentence and subject were included as random effects. Including condition in the model as a fixed effect significantly improved model fit, $\chi^2$ (4) = 9.98, $p$ = 0.041, suggesting that alignment of tokens by speaker intelligibility differentially facilitated generalization of accent learning.[5] Contrast analyses indicated that transcription accuracy in the LHLH condition was reliably higher than in the HHLL condition, $b$ = .40, $p$ = .046. All other comparisons were not significant (all $b$s < .33, all $p$s > .155).

That test performance in the HHLL condition was numerically worse than performance in the other conditions and statistically worse than in the LHLH condition contradicts our prediction that that listeners would achieve highest mean transcription accuracy at test when trained with high- rather than low-intelligibility speakers first. One possibility for this lack of a progressive alignment effect might be that the inclusion of four talkers might have been too few to allow generalizable learning to occur. A related possibility is that the difference

[5]Including condition in the linear model as a fixed effect did *not* significantly improve model fit, $\chi^2$ (4) = 4.17, $p$ = 0.384. A one-way ANOVA comparing transcription accuracy at test across the conditions also showed that transcription accuracy did not reliably differ, $F$(4, 146) = 1.251, $p$ = .193, *partial* $\eta^2$ = .041, suggesting that the results of the logistic mixed model should be interpreted with caution.

between the high and low intelligibility speakers was not optimal for a progressive alignment comparison process to be necessary or beneficial. That is, relative differences between speakers' intelligibility may have been too large or small to effectively scaffold learning of the accent category (Gentner, 1983; Genter & Medina, 1999). Despite the significant difference between LHLH and HHLL conditions, that none of the orderings by speaker intelligibility yielded test performance that differed from that in the Control condition provides evidence that grouping tokens by speaker does not allow listeners to identify systematic variation due to accent.

## General Discussion

The current results suggest that the organization of training tokens influences perceptual learning of an accent category. With amount of variation held constant across training conditions, results showed that relative to trial-by-trial variation in either linguistic content or speaker identity, variation in *both* factors across training trials yielded more accurate transcription of novel tokens spoken by unfamiliar accented speakers. These findings provide novel insight into the process by which listeners update their speech representations to enable generalizable accent learning. Beyond the importance of exposure to sufficient variation during training for generalization to occur, perceptual learning involves the opportunity to align relevant sources of variation across instances. Existing exemplar-based theories of spoken language perception do not account for the role of training structure in perceptual learning. Thus, the current findings qualify the extent to which these models represent the process by which listeners adapt to variation in the speech signal.

In two experiments, English-speaking listeners completed training sessions in which they heard Spanish-accented speech produced by multiple non-native speakers. In Experiment 1, we assessed whether training structure would influence learning of sentence-length utterances. Listeners transcribed sentences with the highest accuracy in the Variable condition, suggesting that training structure impacted perceptual learning. Results from Experiment 1b suggest that this facilitative effect of the Variable training cannot be solely attributed to similarities in task structure between training and test. Experiment 2 assessed the extent to which the progressive alignment of training tokens by a single variable, speaker intelligibility, would facilitate speaker-independent accent learning. Listeners' test performance across alignment types did not differ from performance in the no-training control, suggesting that alignment of tokens by speaker intelligibility did not affect perceptual learning of accented speech.

Overall, findings from the current study suggest that training structure affects the robustness of generalization during accent learning. Specifically, trial-by-trial variation on two dimensions, speaker's voice and linguistic content, appeared to significantly facilitate learning of systematic, speaker-independent accent regularities and did so even when the amount of variation in the training materials was held constant. These findings are consistent with previous work suggesting a facilitative effect of high-variation training on perceptual learning of accented speech (e.g., Bradlow & Bent, 2008; Iverson, Hazan, & Bannister, 2005; Lively et al., 1993; Sidaras et al., 2009; Barcroft & Sommers, 2005; Sumner, 2011). Bradlow and Bent (2008) found that relative to participants who only heard one speaker

during training, those who heard sentences spoken by multiple speakers during training transcribed sentences spoken by a novel accented speaker at test more accurately, suggesting that speaker variation facilitates speaker-independent learning of accent. Similarly, Lively et al. (1993) found that whereas listeners who completed training with the /r/-/l/ contrast produced by multiple speakers generalized their learning to novel tokens and speakers, listeners who heard one speaker at training generalized learning only to novel tokens produced by the same speaker but not to unfamiliar speakers. Taken together, these results suggest that training with tokens that are highly variable across multiple dimensions encourages identification of speaker-independent systematicities in accented speech.

Going beyond previous findings, the results of the current study suggest that generalization under high-variation exposure conditions depends on how stimuli are ordered during training. When the same set of stimuli are blocked by speaker or by sentence, there is no evidence of cross-speaker generalized learning. When the structure of the training phase maximizes trial-to-trial variation across both dimensions, however, cross-speaker learning is robust. Thus, trial-to-trial variation in speaker and sentence content promotes speaker-independent learning of accent. Across experiments, listeners consistently showed better performance when given the opportunity to compare and contrast variation across relevant source dimensions. That there was no evidence of speaker-independent learning in the Speaker and Sentence conditions underscores the significance of comparing across multiple sources of variation in order to achieve accent learning that generalizes to novel speakers and novel linguistic content. Previous findings suggest that speaker variation promotes speaker-independent of foreign-accented speech (e.g., Bradlow & Bent, 2008; Sidaras et al., 2009). However, the current finding that neither speaker- or sentence-blocked training yielded learning suggests that multi-speaker exposure might be necessary but not sufficient for robust speaker-independent learning of accent. Crucially, learning occurs when listeners are provided with the opportunity to compare across multiple sources of variation during exposure.

The effect of training structure in the current study is consistent with literature suggesting that comparing tokens can facilitate category membership learning more generally (Gentner & Namy, 1999; Higgins & Ross, 2011; Kang & Pashler, 2012; Kornell & Bjork, 2008; Namy & Gentner, 2002). The current findings also align with work showing a benefit for spaced versus massed exposure on motor- and verbal-learning tasks (e.g., Schmidt & Bjork, 1992). Trial-to-trial comparison of tokens spoken by different speakers may facilitate the disambiguation of variation due to individual speaker's idiosyncratic pronunciations from speaker-independent regularities in accent. However, given findings by Perrachione et al. (2011) that trial-by-trial variation in a vocabulary learning task benefitted only participants with stronger rather than weaker auditory abilities, future work might consider individual differences when evaluating performance in perceptual learning paradigms.

The ordering of training stimuli may also provide a "desirable difficulty" (Bjork, 1994) that leads to more elaborative processing (e.g., Alter et al., 2007). Variable training included trial-to-trial variation along multiple dimensions and may have been more difficult than the other training exposure conditions. Given this difficulty in processing during Variable training, one possibility is that listeners adopted more effortful strategies to extract speaker-

independent accent sytematicities. Similarly, error-based implicit learning accounts (e.g., Fine & Jaeger, 2013) also suggest a role of training structure in learning outcomes. According to such views, learning elicits prediction errors such that expectations generated from recently processed information affect predictions about upcoming material. Error-based models have been discussed in the context of syntactic acquisition such that learners' representations update to account for erroneous predictions about correct syntactic form (e.g., Chang, Dell, & Bock, 2006). Larger prediction errors (surprisal) increase processing difficulty, which then enhances the learning of correct syntactic forms. Relative to other training conditions, Variable training may induce greater levels of surprisal (e.g., greater trial to trial predictive error), as each trial presents stimuli that is different in both speaker identity and lexical content. This surprisal may enhance the flexibility in listeners' representation of accented speech and facilitate generalization of learning to novel talkers and lexical items.

Not mutually exclusive from this possibility is that training structure may affect the extent to which listeners weigh particular speaker and accent characteristics (Francis & Nusbaum, 2002; Iverson & Kuhl, 1995). Relative to the structure of training in the other conditions, Variable training provides the richest source of information as both speaker identity and linguistic content change with each trial. Optimal test performance for participants in the Variable condition in the current study can potentially be attributed to reweighting acoustic cues from one trial to the next that help to highlight systematic variation due to accent and divert attention from idiosyncratic pronunciations unique to each speaker. Although findings from the current study do not disambiguate among these possible mechanisms, they do suggest that training structure plays an important role in achieving perceptual learning that generalizes to novel speech tokens.

The results of Experiment 2 suggest that progressive alignment of training tokens by speaker intelligibility did not facilitate speaker-independent accent learning, at least in this training context. However, these findings are not entirely inconsistent with evidence suggesting that alignment of training tokens on a *specific* dimension facilitates perceptual learning of speech (Church et al., 2013; Iverson et al., 2005; Sumner, 2011). Sumner's (2011) results, for example, suggest that aligning training tokens by a specific acoustic property such as VOT may be more likely to elicit perceptual learning. Similarly, Church et al. (2013) showed that participants' ability to discriminate between bird song rates varied as a function of training structure. During training, those who heard the original bird song rate paired with stimuli ordered from most to least different from the original (easy-to-hard or progressive training), performed better on a discrimination task with both familiar and novel stimuli than those who completed anti-progressive, or random training. In the current study, speaker intelligibility is defined by the mean transcription accuracy of each speaker's utterances and thus implies using more global or diverse attributes to categorize low and high intelligibility speakers. Indeed, alignment by speaker intelligibility necessitates grouping by a host of acoustic-phonetic dimensions, including characteristic vowel space and segmental timing relations (Bradlow et al., 1996). The finding that test performance in Speaker and Sentence conditions in Experiment 1 did not differ from that in the Control condition, along with little evidence for learning in Experiment 2, suggests alignment along a single relevant dimension (i.e., speaker or linguistic item), or progressive alignment along a global dimension such as intelligibility may not facilitate learning. Rather, aligning tokens by more specific acoustic-

phonetic features may yield learning that generalizes to novel speakers and lexical items. It remains for future research to determine whether particular acoustic-phonetic characteristics of foreign accented speech can be aligned to elicit optimal perceptual learning.

One alternative explanation for the higher test performance in the training conditions relative to the no-training control is that training increases task familiarity. However, findings from perceptual learning studies that include training with tokens spoken by native English speakers (e.g., Sidaras et al., 2009) indicate no reliable performance differences between native English training and no training for both word and sentence stimuli. In addition, findings from Experiment 1b eliminated task familiarity as a potential explanation for better performance at test in the Variable condition, as listeners who completed Spanish Variable training outperformed those who completed English Variable training or no training. Thus, differences between conditions in the current study cannot solely be attributed to familiarity with the task itself.

A second potential explanation for the differential test performance across conditions is that the conditions varied in the temporal distance between repetitions of tokens and hence perhaps, in possible memory constraints. A listener in the Variable condition, for example, could hear approximately 36 sentences during training before hearing any given sentence for a second time, whereas a listener in the Sentence condition would hear a given sentence repeated in the very next trial. Perhaps listeners derived some learning benefit from the delay in repetition of sentence content. However, since repetitions of each sentence were of similar temporal distances in both the Variable and Speaker conditions, this difference in temporal distance of stimuli repetitions cannot fully account for the optimal performance in the Variable versus Speaker conditions. Taken together, the findings suggest that it is primarily the organization of the training materials or training structure itself that influenced perceptual learning.

The finding that training structure affects perceptual learning provides clear constraints on models of speech perception to account for the role of exposure structure in speech perception. Although the results of the present study align in part with exemplar-based models of speech perception (e.g., Goldinger, 1996; 1998), which claim that highly detailed surface characteristics of spoken utterances, such as speaker’s voice, speaking rate, intonation contour, and $F_0$, are encoded in listeners' representations of spoken utterances, currently unaccounted for in exemplar-based models is a role for training structure in perceptual learning. This class of model posits that exposure to highly variable speech tokens yields more widely distributed representations that then enable generalizable learning. However, exposure to similar distributions of variable tokens should lead to equivalent generalization of learning. That training structure in the current study influenced perceptual learning when the number and type of training tokens were held constant suggests that exemplar-based models, and indeed models of speech processing more generally, will need to be expanded to accommodate the influence of structured exposure to variation. For example, existing exemplar models of speech perception might incorporate a level of processing that encodes the sequencing or time course of listeners’ exposure to spoken exemplars. Accounting for training structure would affect which episodic traces are activated and either facilitate or interfere with the integration of newly encountered

exemplars into memory, and ultimately, the accuracy of word recognition and perceptual learning processes.

One computational model that may account for the current findings is Kleinschmidt and Jaeger's (2015) ideal adapter framework. According to this view, optimal perceptual learning varies as a function of the statistical distribution of previously experienced percepts, the listener's beliefs or knowledge about the statistics of the relevant categories, and the listener's belief that there is a need to adapt. Perceptual learning of non-native accented speech occurs when listeners experience a distribution of pronunciations characteristic of a given accent. This prior experience then provides the foundation for generative models that combine bottom-up acoustic information and top-down expectations to adapt to the accented utterance. Implied in the ideal adapter framework is the idea that the *organization* of experience with accented speech influences the type of generative model that the listener implements for understanding accented speech. Thus, Variable training, with meaningful trial-by-trial variation in both speaker identity and linguistic content, may give rise to a particularly representative and diagnostic set of expectations that allow the listener to accommodate novel accented pronunciations produced by novel speakers.

### Conclusions

The results of the present study provide support for the significance of training structure in perceptual learning outcomes. Whereas previous work has found evidence for the facilitative effect of high-variation training on accent learning, the results of the current study suggest that beyond the type of variation among training tokens, the organization of tokens affects the robustness of learning as well. Overall, these results suggest that identification of systematic, speaker-independent regularities can be facilitated by manipulating the structure of exposure to variation during the perceptual learning of accented speech. Listeners are not only encoding the range and type of variation inherent in the acoustic speech signal but also engaging in a perceptual process that may involve online comparison of similarities and differences across tokens to identify accent- from speaker-dependent sources of variation. Given that extant theories of spoken language processing currently do not account for such training structure effects, the present findings suggest the need to incorporate the consequences of exposure organization for perceptual learning.

## Acknowledgments

This research was supported in part by Research Grant R01 DC 008108 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. Portions of this work were presented at the International Meeting of the Acoustical Society of America and the Annual Meeting of the Psychonomic Society. We thank the editor and reviewers for their insightful comments on an earlier version of this manuscript.

## References

Allen JS, Miller JL. Listener sensitivity to individual talker differences in voice-onset-time. The Journal of the Acoustical Society of America. 2004; 115(6):3171–3183. [PubMed: 15237841]

Alter AL, Oppenheimer DM, Epley N, Eyre RN. Overcoming intuition: metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General. 2007; 136(4):569–576. [PubMed: 17999571]

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59(4):390–412.

Barr DJ, Levy R, Scheepers C, Tilly HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language. 2013; 68:255–278.

Barcroft J, Sommers MS. Effects of acoustic variability on second language vocabulary learning. Studies in Second Language Acquisition. 2005; 27(3):387–414.

Bates, D.; Maechler, M.; Bolker, B.; Walker, S. lme4: Linear mixed-effects models using Eigen and S4. 2014. R package version 1.1-7, http://CRAN.R-project.org/package=lme4

Bjork, RA. Memory and metamemory considerations in the training of human beings. In: Metcalfe, J.; Shimamura, A., editors. Metacognition: Knowing About Knowing. Cambridge, MA: MIT Press; 1994. p. 185-205.

Bond ZS, Moore TJ. A note on the acoustic-phonetic characteristics of inadvertently clear speech. Speech Communication. 1994; 14(4):325–337.

Bradlow AR, Bent T. Perceptual adaptation to non-native speech. Cognition. 2008; 106(2):707–729. [PubMed: 17532315]

Bradlow AR, Pisoni DB. Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. The Journal of the Acoustical Society of America. 1999; 106(4): 2074–2085. [PubMed: 10530030]

Bradlow AR, Torretta GM, Pisoni DB. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication. 1996; 20(3):255–272. [PubMed: 21461127]

Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin. 2006; 132(3):354–380. [PubMed: 16719566]

Chang F, Dell GS, Bock K. Becoming syntactic. Psychological Review. 2006; 113(2):234–272. [PubMed: 16637761]

Church BA, Mercado E III, Wisniewski MG, Liu EH. Temporal dynamics in auditory perceptual learning: Impact of sequencing and incidental learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2013; 39(1):270–276.

Clarke CM, Garrett MF. Rapid adaptation to foreign-accented English. The Journal of the Acoustical Society of America. 2004; 116(6):3647–3658. [PubMed: 15658715]

Clutterbuck R, Johnston RA. Demonstrating how unfamiliar faces become familiar using a face matching task. European Journal of Cognitive Psychology. 2005; 17(1):97–116.

Creel SC, Aslin RN, Tanenhaus MK. Heeding the voice of experience: The role of talker variation in lexical access. Cognition. 2008; 106(2):633–664. [PubMed: 17507006]

Cutler A. The abstract representations in speech processing. The Quarterly Journal of Experimental Psychology. 2008; 61(11):1601–1619. [PubMed: 18671170]

Cutler A, Eisner F, McQueen JM, Norris D. How abstract phonemic categories are necessary for coping with speaker-related variation. Laboratory Phonology. 2010; 10:91–111.

Dahan D, Drucker SJ, Scarborough RA. Talker adaptation in speech perception: Adjusting the signal or the representations? Cognition. 2008; 108(3):710–718. [PubMed: 18653175]

Dwyer DM, Vladeanu M. Perceptual learning in face processing: Comparison facilitates face recognition. The Quarterly Journal of Experimental Psychology. 2009; 62(10):2055–2067. [PubMed: 19235097]

Eisner F, McQueen JM. The specificity of perceptual learning in speech processing. Perception and Psychophysics. 2005; 67(2):224–238. [PubMed: 15971687]

Fine AB, Jaeger TF. Evidence for implicit learning in syntactic comprehension. Cognitive Science. 2013; 37(3):578–591. [PubMed: 23363004]

Flege JE, Bohn OS, Jang S. Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics. 1997; 25(4):437–470.

Francis AL, Nusbaum HC. Selective attention and the acquisition of new phonetic categories. Journal of Experimental Psychology: Human Perception and Performance. 2002; 28(2):349–366. [PubMed: 11999859]

Gentner D. Structure-mapping: A theoretical framework for analogy. Cognitive Science. 1983; 7(2): 155–170.

Gentner D, Markman AB. Structural alignment in comparison: No difference without similarity. Psychological Science. 1994; 5(3):152–158.

Gentner D, Medina J. Similarity and the development of rules. Cognition. 1998; 65(2):263–297. [PubMed: 9557385]

Gentner D, Namy LL. Comparison in the development of categories. Cognitive Development. 1999; 14(4):487–513.

Glenberg AM. Component-levels theory of the effects of spacing of repetitions on recall and recognition. Memory and Cognition. 1979; 7(2):95–112. [PubMed: 459836]

Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22(5): 1166–1183.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105(2):251–279. [PubMed: 9577239]

Greenspan SL, Nusbaum HC, Pisoni DB. Perceptual learning of synthetic speech produced by rule. Journal of Experimental Psychology: Learning, Memory & Cognition. 1988; 14(3):421–433.

Hazan V, Simpson A. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. Speech Communication. 1998; 24(3):211–226.

Higgins, EJ.; Ross, BH. Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society; 2011. Comparisons in category learning: How best to compare for what; p. 1388-1393.

Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biometrical Journal. 2008; 50(3):346–363. [PubMed: 18481363]

Iverson P, Hazan V, Bannister K. Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. The Journal of the Acoustical Society of America. 2005; 118(5):3267–3278. [PubMed: 16334698]

Iverson P, Kuhl PK. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America. 1995; 97(1):553–562. [PubMed: 7860832]

Jaeger TF. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language. 2008; 59:434–446. [PubMed: 19884961]

Kang SH, Pashler H. Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. Applied Cognitive Psychology. 2012; 26(1):97–103.

Kleinschmidt D, Jaeger TF. Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. Psychological Review. 2015; 122(2):148–203. [PubMed: 25844873]

Kornell N, Bjork RA. Learning concepts and categories: Is spacing the “enemy of induction”? Psychological Science. 2008; 19(6):585–592. [PubMed: 18578849]

Kraljic T, Samuel AG. Perceptual learning for speech: Is there a return to normal? Cognitive Psychology. 2005; 51(2):141–178. [PubMed: 16095588]

Kraljic T, Samuel AG. Generalization in perceptual learning for speech. Psychonomic Bulletin and Review. 2006; 13(2):262–268. [PubMed: 16892992]

Kraljic T, Samuel AG, Brennan SE. First impressions and last resorts: How listeners adjust to speaker variability. Psychological Science. 2008; 19(4):332–338. [PubMed: 18399885]

Ladefoged, P. UCLA Working Papers in Linguistics. Vol. 73. Los Angeles, CA: University of California; 1989. Representing phonetic structure; p. 1-79.

Ladefoged P, Broadbent DE. Information conveyed by vowels. The Journal of the Acoustical Society of America. 1957; 29(1):98–104.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74(6):431–461. [PubMed: 4170865]

Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. The Journal of the Acoustical Society of America. 1993; 94(3):1242–1255. [PubMed: 8408964]

Locker L, Hoffman L, Bovaird JA. On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. Behavior Research Methods. 2007; 39(4):723–730. [PubMed: 18183884]

McQueen JM, Cutler A, Norris D. Phonological abstraction in the mental lexicon. Cognitive Science. 2006; 30(6):1113–1126. [PubMed: 21702849]

Miller JD. Auditory-perceptual interpretation of the vowel. The Journal of the Acoustical Society of America. 1989; 85(5):2114–2134. [PubMed: 2659639]

Munro MJ. Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. Language and Speech. 1993; 36(1):39–66. [PubMed: 8345772]

Namy LL, Gentner D. Making a silk purse out of two sow's ears: young children's use of comparison in category learning. Journal of Experimental Psychology: General. 2002; 131(1):5–15. [PubMed: 11900103]

Norris D, McQueen JM, Cutler A. Perceptual learning in speech. Cognitive Psychology. 2003; 47(2): 204–238. [PubMed: 12948518]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception and Psychophysics. 1998; 60(3):355–376. [PubMed: 9599989]

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychological Science. 1994; 5(1):42–46. [PubMed: 21526138]

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19(2):309–328.

Perrachione TK, Lee J, Ha LY, Wong PC. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. The Journal of the Acoustical Society of America. 2011; 130(1):461–472. [PubMed: 21786912]

Peterson GE, Barney HL. Control methods used in a study of the vowels. The Journal of the Acoustical Society of America. 1952; 24(2):175–184.

R Core Team. R Foundation for Statistical Computing. Vienna, Austria: 2014. R: A language and environment for statistical computing. URL http://www.R-project.org/

Raaijmakers JG. A further look at the" language-as-fixed-effect fallacy". Canadian Journal of Experimental Psychology. 2003; 57(3):141–151. [PubMed: 14596473]

Rothauser EH, Chapman WD, Guttman N, Nordby KS, Silbiger HR, Urbanek GE, Weinstock M. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 1969; 17(3):225–246.

Schmidt RA, Bjork RA. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. Psychological Science. 1992; 3(4):207–217.

Schneider, W.; Eschman, A.; Zuccolotto, A. E-Prime: User's guide. Psychology Software Incorporated; 2002.

Sidaras SK, Alexander JED, Nygaard LC. Perceptual learning of systematic variation in Spanish-accented speech. The Journal of the Acoustical Society of America. 2009; 125(5):3306–3316. [PubMed: 19425672]

Sjerps MJ, McQueen JM. The bounds on flexibility in speech perception. Journal of Experimental Psychology: Human Perception and Performance. 2010; 36(1):195–211. [PubMed: 20121304]

Sumner M. The role of variation in the perception of accented speech. Cognition. 2011; 119(1):131–136. [PubMed: 21144500]

Trude AM, Brown-Schmidt S. Talker-specific perceptual adaptation during online speech perception. Language and Cognitive Processes. 2012; 27(7–8):979–1001.

Yonan CA, Sommers MS. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychology and Aging. 2000; 15(1):88–99. [PubMed: 10755292]

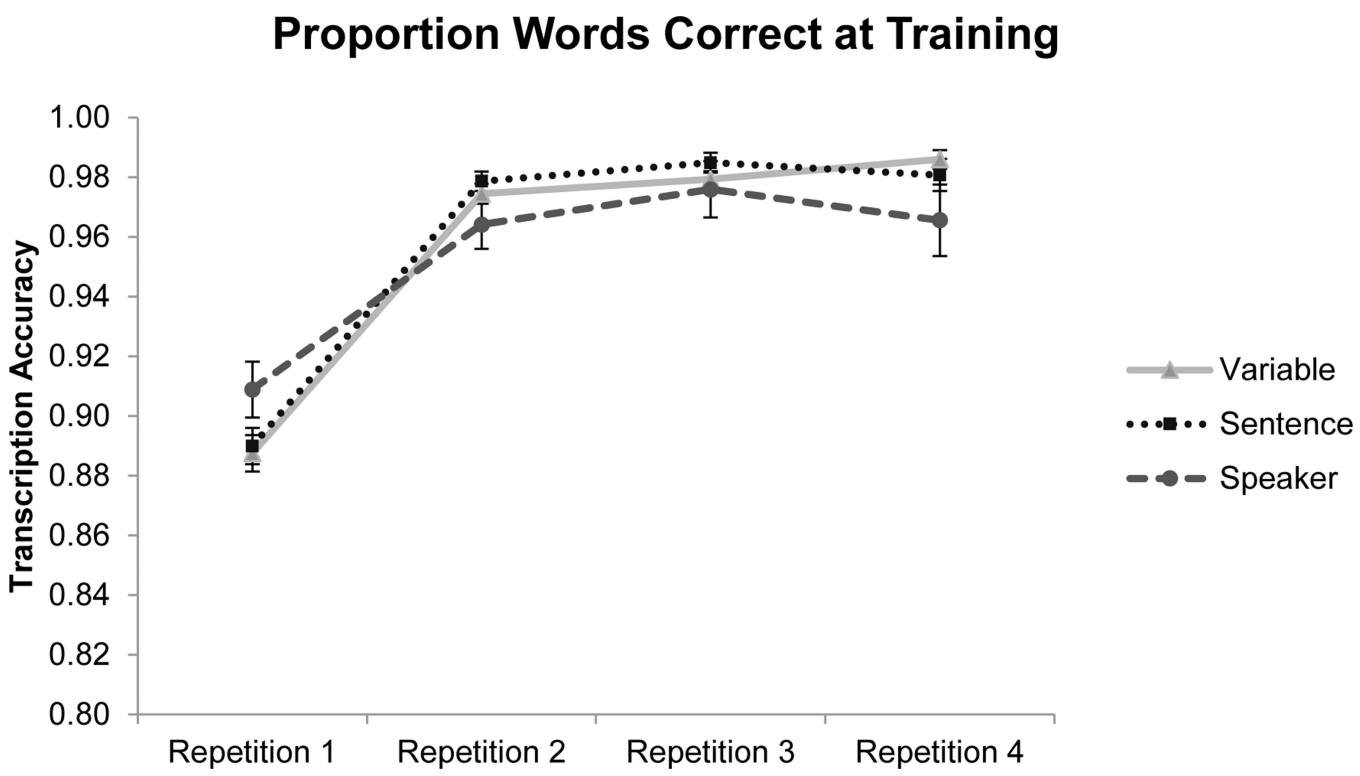


**Figure 1.**

Transcription accuracy during training for each sentence repetition in the three training conditions in Experiment 1. Across conditions, listeners' transcription accuracy was significantly higher for the last versus the first repetition, suggesting that participants engaged in perceptual learning for all conditions in the training phase. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

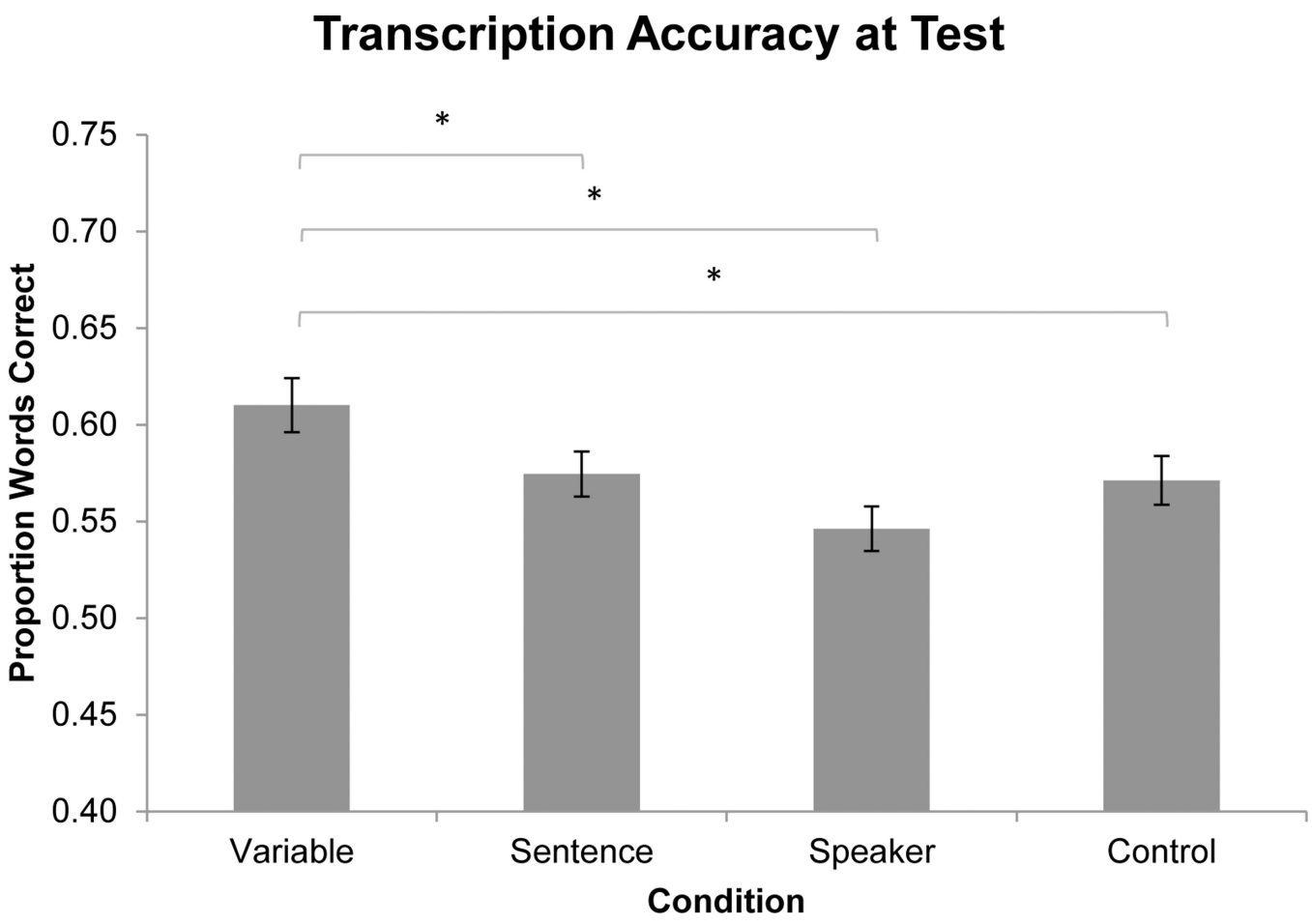


**Figure 2.**

Transcription accuracy at test as a function of condition in Experiment 1. Transcription accuracy was significantly higher in the Variable condition than in the Control, Sentence, and Speaker conditions. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

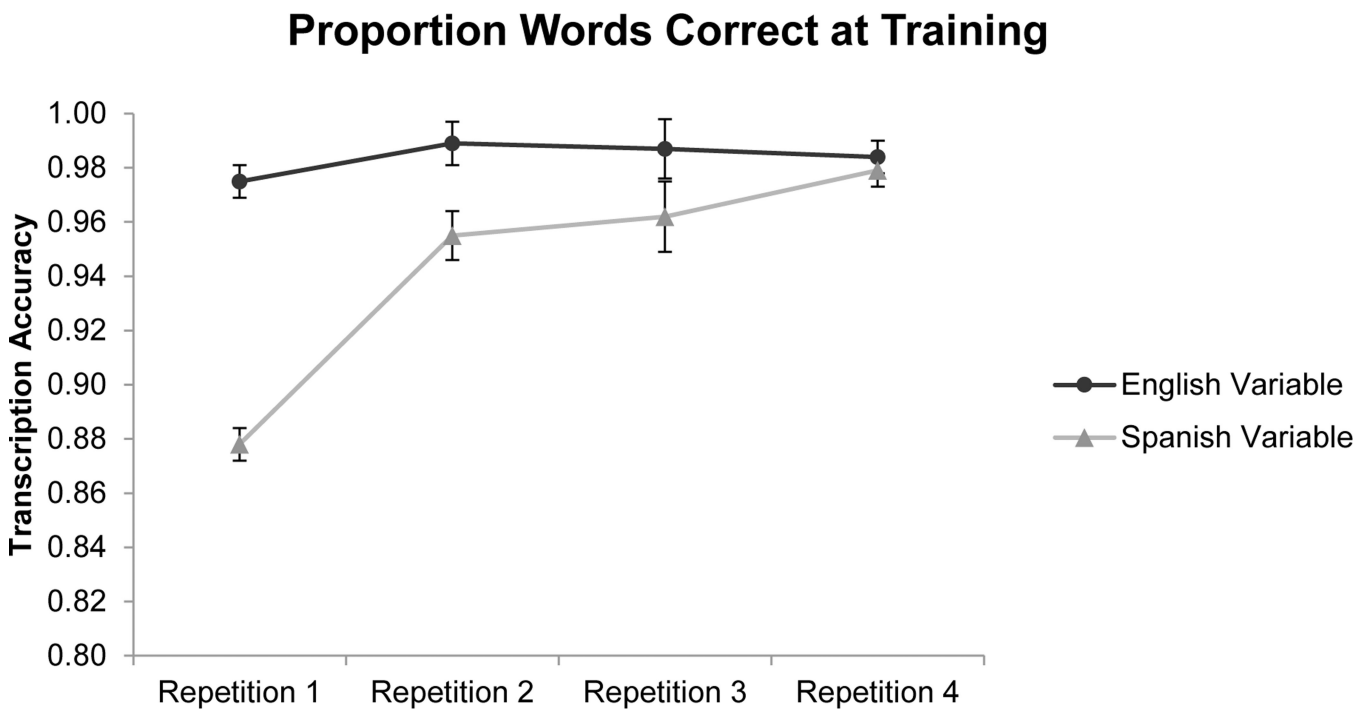


**Figure 3.**

Transcription accuracy during training for each sentence repetition in the three training conditions in Experiment 1b. By repetition 4, transcription accuracy is near ceiling, suggesting that participants engaged in perceptual learning for both conditions in the training phase. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

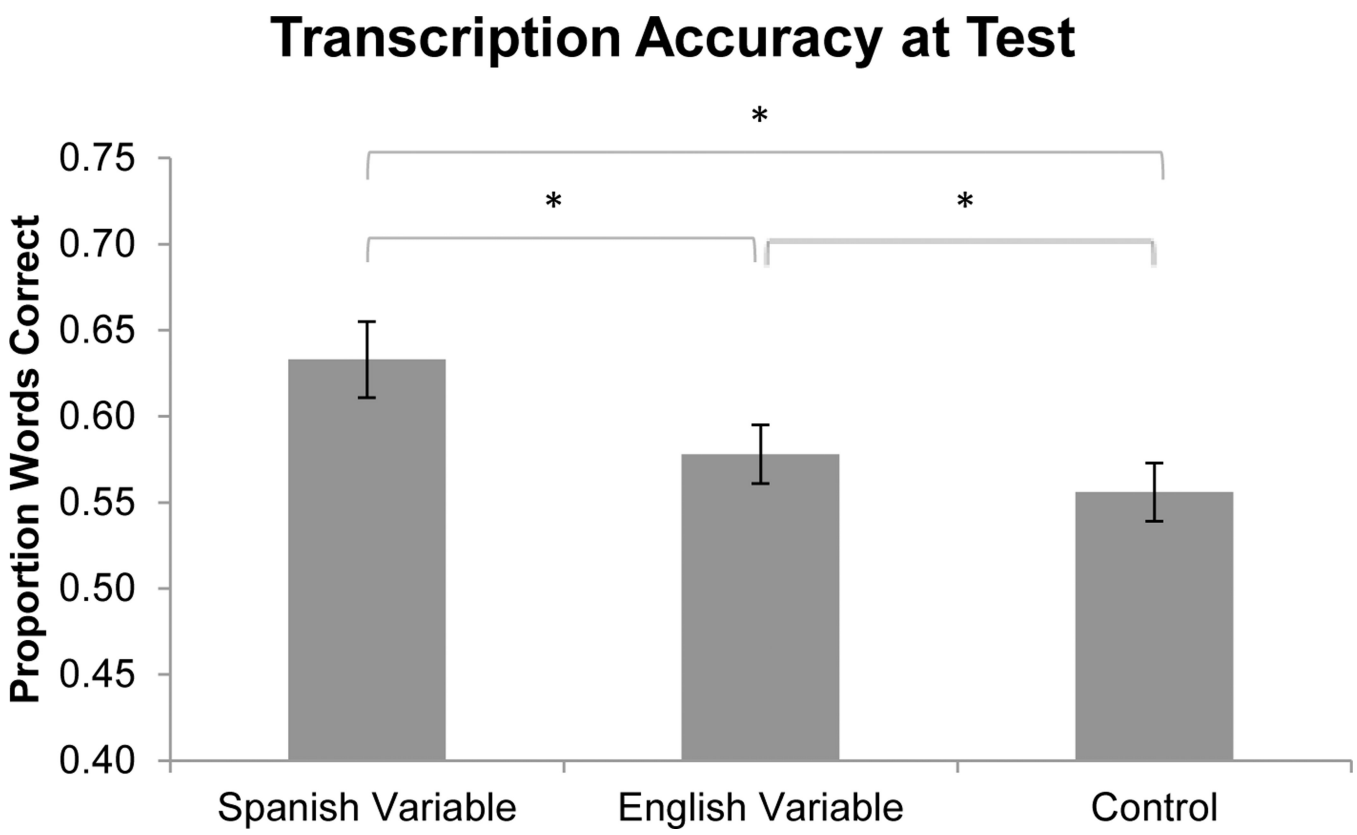


**Figure 4.**

Transcription accuracy at test as a function of condition in Experiment 1b. Critically, transcription accuracy was significantly higher in the Spanish Variable condition than in the English Variable and Control conditions. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

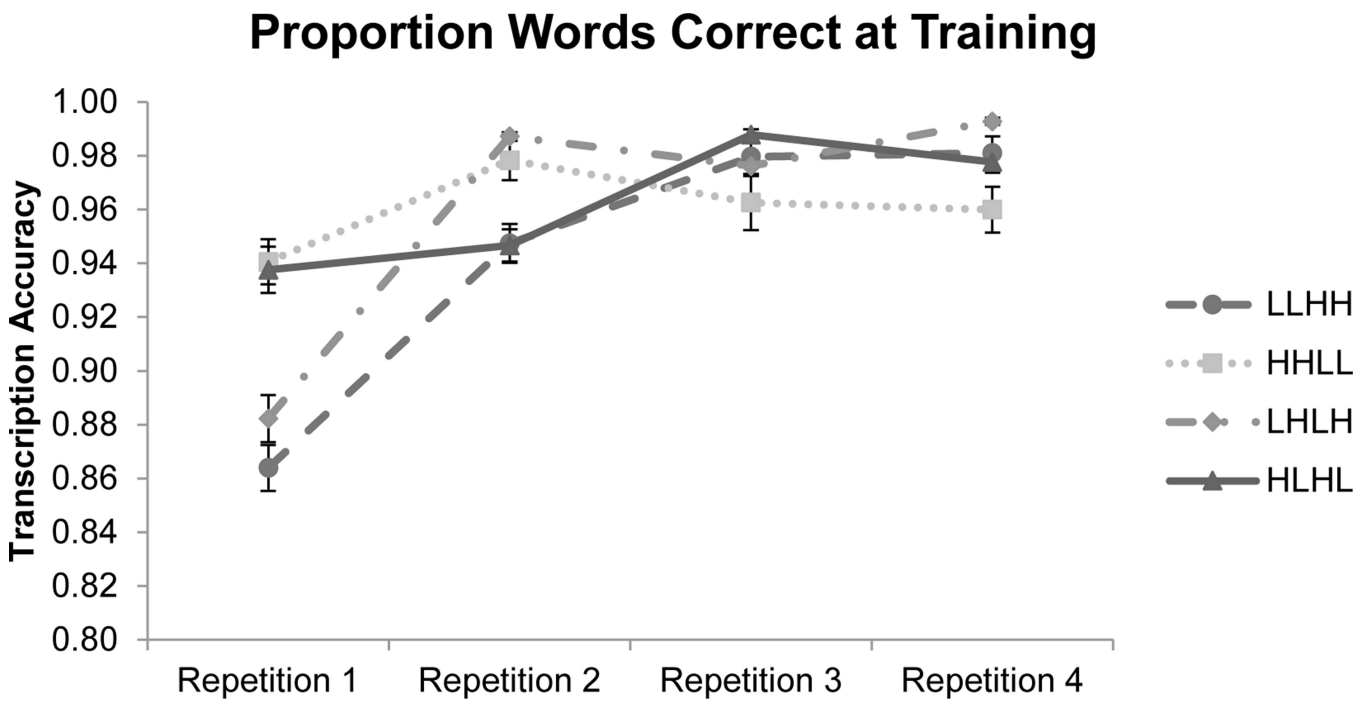


**Figure 5.**

Transcription accuracy during training for each sentence repetition in the four training conditions in Experiment 2. Across conditions, listeners' transcription accuracy was significantly higher for the last versus the first repetition, suggesting that participants engaged in perceptual learning for all conditions in the training phase. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

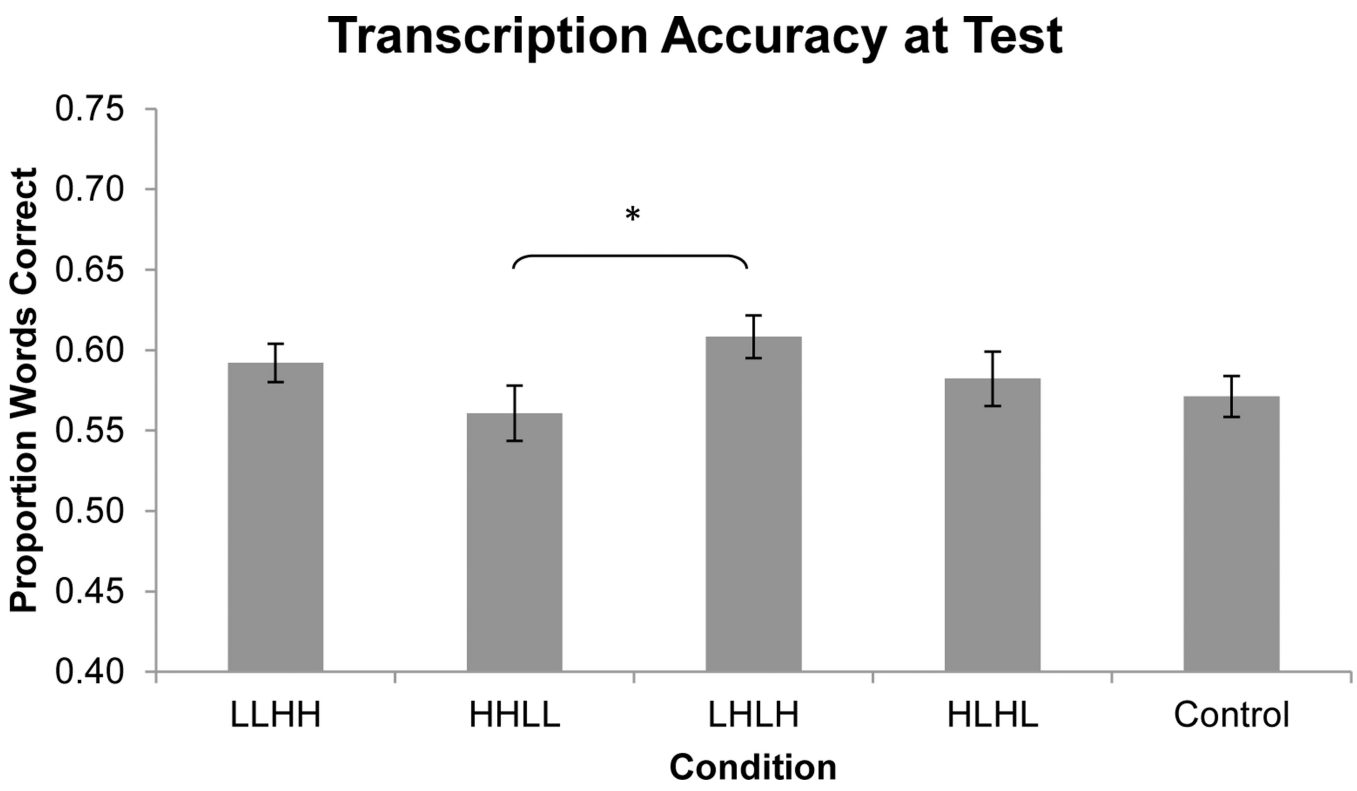


**Figure 6.**

Transcription accuracy at test as a function of condition in Experiment 2. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

## Table 1

Accentedness and intelligibility for Spanish-accented speakers

| Speaker Group | Gender | Mean Accentedness Ratings | Mean Intelligibility Scores (Sentences) (%) |
|---|---|---|---|
| **Group 1** | Female | 5.59 | 75.6 |
| | Female | 3.10 | 89.8 |
| | Male | 4.77 | 65.9 |
| | Male | 2.83 | 90.5 |
| **Group 2** | Female | 6.17 | 74.6 |
| | Female | 4.31 | 85.5 |
| | Male | 4.75 | 89.0 |
| | Male | 3.55 | 81.8 |

*Note*. Listeners rated the accentedness of each sentence on a 7-point Likert-type scale, from 1 = "not accented" to 7 = “very accented”.