Attention, Perception, & Psychophysics (2019) 81:981–1005
https://doi.org/10.3758/s13414-019-01725-4

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Interactions between speech perception and production during learning of novel phonemic categories

Melissa Michaud Baese-Berk[1]

Published online: 11 April 2019

Abstract
A successful language learner must be able to perceive and produce novel sounds in their second language. However, the relationship between learning in perception and production is unclear. Some studies show correlations between the two modalities; however, other studies have not shown such correlations. In the present study, I examine learning in perception and production after training in a distributional learning paradigm. Training modality is manipulated, while testing modality remained constant. Overall, participants showed substantial learning in the modality in which they were trained; however, learning across modalities shows a more complex pattern. Although individuals trained in perception improved in production, individuals trained in production did not show substantial learning in perception. That is, production during training disrupted perceptual learning. Further, correlations between learning in the two modalities were not strong. Several possible explanations for the pattern of results are explored, including a close examination of the role of production variability, and the results are explained using a paradigm appealing to shared cognitive resources. The article concludes with a discussion of the implications of these results for theories of second-language learning, speech perception, and production.

Keywords Speech perception · Speech production · Perceptual learning

In order to successfully communicate in a language, an individual must be able to both perceive and produce that language. Most current theories of speech perception or production assume a relatively straightforward relationship between the two modalities. That is, the two modalities are assumed to share representations and processes. In many models of speech perception, single word perception begins with auditory processing, and these sounds are then mapped onto phonetic and phonological representations, lexical representations, and semantic representations. Production is often described as being the nearly same process in reverse, beginning with accessing a semantic representation, a lexical representation, and then sound structure, before a word is produced using articulators. An abundance of recent work has shifted the focus from perception and production in isolation to the relationship between the two modalities. However, several recent studies have suggested that the relationship between perception and production may not be as straightforward as commonly assumed. The present studies examine the interaction between speech perception and production in a specific case: learning nonnative speech sound categories using a distributional learning paradigm. I present evidence that the relationship between perception and production at the earliest stages of nonnative speech category learning in each modality is complex and suggest that these data support a reformulation of current theories of both perception and production to account for the complexity of this relationship.

✉ Melissa Michaud Baese-Berk
mbaesebe@uoregon.edu

[1] Department of Linguistics, 1290 University of Oregon, Eugene, OR 97403, USA

Relationship between perception and production

As stated above, the relationship between perception and production is the topic of an ongoing debate in the speech community.[1] Although it is clear that perception and production must interact in the systems of proficient, adult speakers of a

[1] This debate is not limited to human speech. Substantial work in birdsong has examined perception and production links post adulthood (see Prather, Okanoya, & Bolhuis, 2017, for a review).

language, it is unclear how this interaction occurs and how best to characterize the nature of the relationship between the two modalities. The strongest views on this matter have been posited by those who suggest that the two modalities are very closely related. For example, both the motor theory of speech production (Libermann, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Libermann, Delattre, & Cooper, 1952; Libermann & Mattingly, 1985, 1989) and direct realist theories (e.g., Best, 1995; Fowler, 1986) posit common representations shared between perception and production during processing based on articulatory properties or motor representations.[2]

These theories can be contrasted with a general auditory account of perception (see Diehl, Lotto, & Holt, 2004, for a review), which suggests that the object of speech perception is an acoustic target rather than an auditory one. Under this account, a variety of configurations are possible for the relationship of perception to production. For example, it is possible that both modalities share the same target; however, unlike direct realism or motor theory, this target would be an acoustic representation rather than a motor representation. This configuration is supported by work examining compensation for motor perturbation during production. In these studies, participants compensate for perturbation to achieve an acoustic target, suggesting that there is an acoustic component to production targets (e.g., Guenther, Hampson, & Johnson, 1998). Additional possible configurations under this account would not rely on identical targets in the two modalities; instead, it is possible that the targets of perception and production are distinct, and the modalities are linked at later stages of processing.

Previous experimental work has demonstrated mixed results, with some studies suggesting a close relationship between the two modalities (e.g., Goldinger, 1998), whereas other work suggests that there are few correlations between performance in the two modalities. In a recent intriguing finding, some work has also suggested that the two modalities may have a more antagonistic relationship (Baese-Berk & Samuel, 2016). Given the large variation in findings, it is important to continue investigations and to ask under what circumstances we may expect to see correlations or lack of correlations in performance between the two modalities, and how determining these circumstances could influence our understanding and theories of the interactions between perception and production. The present study provides an extension of the findings of Baese-Berk and Samuel (2016) and also suggests that production may, in some circumstances, disrupt perceptual learning, even when participants learn in production. Below, I review a selection of studies that investigate the relationship between perception and production to demonstrate the complex nature of previous findings.

Several studies have demonstrated some evidence for a close link between perception and production. Using imitation and shadowing (i.e., direct repetition without instruction for imitation), Goldinger and colleagues have demonstrated that tokens produced after exposure to a perceptual target are judged to be perceptually more similar to the target word than baseline productions (i.e., productions made before any exposure to the target speaker; Goldinger, 1998; Goldinger & Azuma, 2004). Shockley, Sabadini, and Fowler (2004) showed that not only are shadowed words judged to be more perceptually similar to target words but that specific acoustic properties of speakers' shadowed tokens shift toward shadowed targets. When shadowing words with lengthened voice onset times (VOTs), speakers produced tokens with lengthened VOTs compared with the VOT of their baseline productions. This also suggests that on a fairly short time scale, fine-grained properties of perception can be transferred to production. Several other studies (e.g., Nye & Fowler, 2003; Vallabha & Tuller, 2004) have further examined the acoustic properties of shadowed speech, demonstrating that in shadowed speech some properties of the perceptual tokens are reflected in production. In a more naturalistic task, Pardo (2006) examined the relationship between speech perception and speech production via phonetic convergence during conversations. Using perceptual similarity ratings by naïve listeners, this data suggest that during a dialogue, speakers alter their speech to be more similar to that of their partner.

However, similar studies suggest that this process of phonetic convergence is modulated by a variety of factors. In an examination of what types of phonetic properties are imitated during shadowing, Mitterer and Ernestus (2008) suggested that only "phonologically relevant" properties are shadowed, and "phonologically irrelevant" properties are not. Specifically, they demonstrated that native Dutch speakers shadow prevoicing generally (compared with short-lag voicing), but the amount of prevoicing is not shadowed. However, using a combination of shadowing and short-term training, Nielsen (2011) showed that individuals shift their productions of VOT (a phonetically irrelevant contrast under Mitterer & Ernestus's definition) to be closer to that of a target voice without any explicit instruction. Babel and colleagues (Babel, 2011; Babel, McGuire, Walters, & Nicholls, 2014) demonstrated that the amount of shadowing is influenced by linguistic factors (e.g., vowel is being shadowed) as well as social factors. Brouwer, Mitterer, and Huettig (2010) examined shadowing of canonical and reduced tokens and found that participants' shadow both types of tokens, but do not shadow the magnitude of difference between canonical and reduced tokens.

These studies suggest that shadowing and accommodation depends on any number of factors, complicating the

[2] See Diehl and Kluender (1989) for alternative accounts to speech perception positing auditory properties as the object of speech perception.

interpretation of these results for understanding the relationship between perception and production more specifically. Further, there are methodological and theoretical considerations in interpreting these findings. Only some of the studies assess similarity of the produced and perceived tokens based on acoustic measurements (Mitterer & Ernestus, 2008; Nielsen, 2011; Shockley et al., 2004; Vallabha & Tuller, 2004); many others rely instead on listeners' perceptual similarity judgments. Although this sort of judgment implies that there are changes in production, it is more difficult to identify any single acoustic property or set of properties to demonstrate that productions are physically more similar to the target tokens. More critically, however, the bulk of these studies have examined aspects of production in the listener's native language, or at least a language they speak proficiently. Therefore, it is difficult to know how general these patterns are, especially during development. Nonnative speech sound learning provides a possible avenue for examination of this relationship.

Speech sound learning

One of the hallmarks of perception of sounds in one's native language is categorical perception, characterized by sharp categorization boundaries between sound categories of a language, and good discrimination across category boundaries found in a language, but not within categories of that language. Many studies have asked how an individual system becomes tuned to the native language, and whether it is possible to retune an individual's system to a nonnative language. At a very early age, infants are able to discriminate relatively well between a wide variety of phonetic contrasts, both native and nonnative; however, they become less sensitive to nonnative contrasts during the first year of life (e.g., Werker & Tees, 1984). By adulthood, listeners are typically insensitive to most contrasts not found in their native language (MacKain, Best, & Strange, 1981). For example, native English listeners are able to categorize tokens from an /r/–/l/ continuum into two categories, a contrast that occurs in their native language. However, Japanese listeners demonstrate poor categorization of those same sounds, because the distinction does not exist in their native language (e.g., Goto, 1971). That is, their perception is reliant on the category structure of their language (e.g., Best, McRoberts, & Sithole, 1988; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Libermann, Harris, Hoffman, & Griffith, 1957; Pegg, Werker, Ferguson, Menn, & Stoel-Gammon, 1992).

Over the past 3 decades, dozens of studies have demonstrated that listeners are able to increase their sensitivity to contrasts that are not found in their native language with training (Strange & Dittman, 1984). In the laboratory, various methods have been used to train nonnative listeners on the perception of novel phonetic contrasts (see Iverson, Hazan, & Bannister, 2005, for a comparison). These investigations have examined a variety of segments including Japanese listeners' perception of English /r/ and /l/ (e.g., Logan, Lively, & Pisoni, 1991), English listeners' perception of a three-way voicing contrast (e.g., McClaskey, Pisoni, & Carrell, 1983; Tremblay, Kraus, & McGee, 1998), English listeners' perception of German vowels (e.g., Kingston, 2003), Spanish and German listeners' perception of English vowels (e.g., Iverson & Evans, 2009), and English listeners' perception of Mandarin tones (e.g., Wang, Spence, Jongman, & Sereno, 1999). In many cases cited above, after relatively brief exposure, listeners are better able to categorize sounds and/or discriminate between categories in a nonnative language. However, it is important to note that there is substantial variability in learner's abilities to learn nonnative speech sounds, including, but not limited to their initial abilities to perceive or produce the contrast (see, e.g., Perrachione, Lee, Ha, & Wong, 2011).

Models of nonnative phoneme learning

In addition to types of training paradigms, other factors may also influence learning of novel categories. For example, Best et al. (1988) demonstrated that native English listeners are quite good at discriminating between Zulu clicks even though the contrast does not exist in English. They suggested that the relationship between sounds in the learner's native language and in the target nonnative language could affect the listener's ability to discriminate. The two predominant models of nonnative and second-language speech perception (the perceptual assimilation model: Best, 1994; Best, McRoberts, & Goodell, 2001; Best, McRoberts, & LaFleur, 1995; Best et al., 1988; PAM-L2: Best & Tyler, 2007; and the speech learning model: Flege, 1995, 1997) make explicit predictions that how listeners will perceive nonnative contrasts and the ease (or difficulty) with which they learn them is directly shaped by the similarity of these contrasts to contrasts in their native language.[3]

Both PAM and SLM make strong claims about the relationship between perception and production. Although PAM itself does not make strong claims regarding production of novel contrasts, the model does posit that speech perception and production share representations. Further, perceptual assimilation under this hypothesis is driven by phonetic similarity of sounds. Because of these general claims, one is able to infer that learning in one modality should be strongly

[3] A third model—second language linguistic perception (L2LP; Leussen & Escudero, 2015)—has emerged as a competitor for these models, making similar explicit predictions about the ease or difficulty with which native sounds will be learned; however, this model is not designed to account for L2 production and will not be discussed in substantial detail here.

correlated to learning in the other modality. Because PAM posits a very close relationship between the two modalities, it is assumed to be the case that learning in each modality will be correlated under this model. SLM makes more explicit claims about the relationship of perception and production during learning. Specifically, it is claimed that perception *leads* production (i.e., should always occur first in terms of learning), and that perception and production become closer to one another over the course of learning.

However, evidence for this procession of learning is limited. Several studies have found evidence that directly contradicts these hypotheses. For example, Sheldon and Strange (1982) demonstrated that production learning can precede perceptual learning. Bradlow, Pisoni, Akahane-Yamada, and Tohkura, (1997) examined whether perceptual training transferred to production learning. At a population level, they demonstrated that learners improved on production of the tokens even without overt production training (for additional evidence of transfer from perceptual training to production learning, see Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Wang, Jongman, & Sereno, 2003). However, the pattern of individual learning is much less clear. Some individual participants show substantial improvement in both perception and production. Others demonstrate improvement in perception alone, with no improvement in production. Still other participants show what are assumed to be floor or ceiling effects. Further, some participants demonstrate improvement in production and do not demonstrate any improvement in perception. This result runs counter to the predictions of PAM and SLM, both of which suggest such improvement should not occur in the absence of perceptual learning.

Of course, perception and production are starkly different tasks in terms of their demands on a learner. When producing, especially when repeating, a learner has increased cognitive demands as compared to perception alone. That is, to repeat a token, the learner must first encode what they have heard, and then retrieve a motor plan that corresponds with their percept to appropriately repeat the token. This increased processing demand could disrupt some aspects of learning. That is, if resources are shared between the two modalities during training, it is possible that perception and production may actually have an antagonistic role during learning, with training in one modality reducing the resources available for learning in the opposite modality.

Some recent evidence suggests that production during training does, in fact, incur a cost to the learner. Baese-Berk and Samuel (2016) examined perceptual learning for native Spanish speakers learning a new sound distinction in Basque. After 2 days of training, naïve participants trained in perception alone demonstrated substantial improvements in their ability to perceive the novel contrast. However, participants who were trained in a paradigm that alternated between perception exposure and production practice did not demonstrate learning in perception. That is, perceptual learning was disrupted by producing tokens during training. Interestingly, participants with more experience with the contrast (i.e., late learners of Basque) demonstrate less disruption to perceptual learning. This study, which is foundational for the current manuscript, is discussed in more detail below.

Previous work has also examined learning in each modality as a result of production-focused training. Hattori (2010) examined the perception and production of /r/ and /l/ by native Japanese speakers. He found that the baseline abilities in perception and production of the contrast was not highly correlated. After training listeners using articulatory, production-oriented training their productions of the contrast improved significantly according to a variety of measures; however, their perception was unchanged after this training (see Tateishi, 2013, for a similar finding). This finding contrasts with that of Leather (1990), who trained Dutch participants on the production of four Mandarin words differing in tone. After training in production, he found that participants generalized this learning to perception. However, the author concedes that this result is not conclusive as only one syllable was used during training and testing. Furthermore, there was no pretest, so it is unclear whether the participants were able to perceive this contrast before training.

Current studies

Even given the substantial body of previous work, the links between perceptual learning and production learning remain unclear. Specifically, although it appears that production can occasionally disrupt perceptual learning, it is unclear whether learning can emerge in production, even for those participants who do not learn in perception. Further, it is unclear how distributional information may differentially influence learning in each modality and whether exposure to clear distributions of tokens may impact the relationship between the two modalities.

In Baese-Berk and Samuel (2016), we demonstrated that producing tokens during training could disrupt perceptual learning. In the present study, I extend these results in two important ways: First, I vary the number of days of training (see Experiment 1 vs. Experiment 2), and second, I investigate production learning in addition to perceptual learning. By providing these two extensions, it is possible to begin to answer the question of under what circumstances we might expect to see a disruption of perceptual learning after production during training. Once we better understand those circumstances, we could provide tests of ways in which this disruption could be alleviated, which has potential real-world consequences, in addition to consequences for our scientific understanding of the relationship between the two modalities.

In the present study, a distributional learning paradigm is used to train participants and explicitly manipulate training modality while holding testing modality stable. That is, participants were trained either in perception alone or in a combination of perception and production practice. All participants, regardless of training modality, are tested in both perception and production. This allows a direct examination of the influence of training modality learning in both perception and production of a nonnative sound category. The results of the studies presented here will give us insight into how distributional information influences learning and will elucidate the relationship between the two modalities.

Method

Experimental stimuli and methods

Stimuli Stimuli in all three experiments were modeled on those used in Maye and Gerken (2000), who demonstrated that participants can learn novel speech sound categories after exposure to a bimodal distribution, but not from a unimodal distribution. All stimuli were resynthesized tokens of syllables spoken by a female native speaker of American English. Three separate synthetic continua were formed, following Maye and Gerken, each with the stop consonant in a different vowel environment (i.e., before /a/, /æ/, and / /). Across continua, voice onset time and steepness of formant transitions were held constant. Vowel durations were equated within and across continua. Each continuum included eight equidistant steps. The syllables were resynthesized from naturally produced tokens of a contrast that English listeners are able to produce, but that English does not use contrastively: prevoiced to a short-lag alveolar stop. To produce a prevoiced stop, a speaker's vocal folds are vibrating during the closure for the stop; critically, voicing begins before the release of the consonant. There is usually minimal disruption in vocal fold vibration following the stop release, unlike an aspirated stop. A short-lag stop has a brief period of aspiration after the stop release, and no voicing during the closure of the stop. For more discussion of realizations of this contrast, see Davidson (2016). Two phonetic cues are used to signal this contrast in the continua. The first is VOT (prevoiced vs. short lag); the second is the formant transitions from the stop consonant to the vowel (steeper for the prevoiced end of the continuum, and shallower for the short lag end).

Synthesis of stimuli In order to create the continua, the voicing and formant transitions of a naturally produced token were covaried using Praat (Boersma & Weenink, 2015). The first two formants were resynthesized for each continuum. The tokens were synthesized using the prevoiced token as a base, so that each subsequent step had a smaller amount of prevoicing and less steep formant transitions. Following Maye and Gerken (2000), 78 ms of naturally occurring prevoicing were included in Token 1. Voice onset time was manipulated by gradually reducing the amount of prevoicing at a step size of approximately 13 ms, such that Token 8 had a positive voice onset time of 13 ms. Vowel formants were resynthesized using Praat's LPC algorithm for the first 60 ms of each vowel. Slope was gradually reduced for each token along the continuum, mirroring slopes in Maye and Gerken. Because each vowel required a different end state, the formant transitions differed slightly across each continuum.

Training paradigm Statistical learning studies have provided a means to examine novel category formation under slightly more naturalistic, though still controlled, laboratory training studies. In the present study, I use a distributional learning paradigm, following Maye and Gerken (2000). This paradigm is an ideal tool for examining the relationship between perception and production. Because participants are trained implicitly, no explicit instruction about the sound categories is needed during training, allowing for a more equal training in perception and production. Below, I report the data from the bimodal training groups, as the unimodal training group did not demonstrate significant learning in perception or production.

Procedure All training and testing took place in a large, single-walled sound booth. Visual stimuli were presented on a computer screen. Audio stimuli were presented over speakers at a comfortable volume for the participant. All tasks were self-paced. Production responses were made using a head-mounted microphone. Responses in perception tasks were made using a button box.

Training The training procedure was an implicit learning paradigm that used pictures to reinforce statistical distributional information given to participants. The procedure for training followed the basic procedure used by Maye and Gerken (2000). In Experiment 1, training took place over 2 consecutive days. In Experiment 2, training occurred over 3 consecutive days. Each day, training was broken into several blocks, with 16 repetitions of the target stimuli. The number of repetitions of each token on the continuum was determined by participant training group.

Figure 1 shows the distribution of tokens in each of the two types of training. Participants in the unimodal training groups received more repetitions of stimuli in the middle of the continuum (i.e., stimuli at Points 4 and 5 on the continuum) and fewer repetitions of those stimuli near the ends of the continuum (i.e., Stimuli 2 and 7), which created a single distribution on the continuum. Participants in the bimodal training groups, on the other hand, received more repetitions of stimuli at two

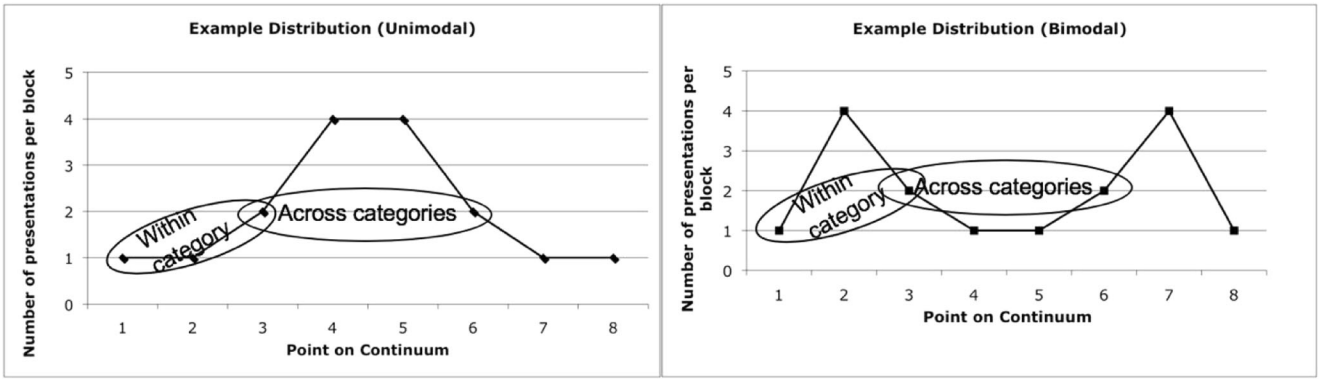

Fig. 1 Distribution of tokens per block for the unimodal (left panel) and bimodal (right panel). Ovals show comparisons that are used in the discrimination test

points along the continuum (i.e., stimuli at Points 2 and 7 on the continuum) and fewer repetitions of the stimuli at the middle of the continuum (i.e., Stimuli 4 and 5), which creates two separate distributions on the continuum. (See Fig. 1 for examples of these distributions.) Participants in the bimodal training group should infer two novel categories, and participants in the unimodal group should infer only one category. Each participant heard 16 experimental tokens from each of three continua, for a total of 48 tokens per block. Participants were exposed to eight training blocks per day for a total of 384 training tokens each day.

All tokens were paired with a picture during presentation. The unimodal group saw one picture per continuum. The continuum for the bimodal group was divided in half, with one picture per half. These pictures reinforced the distributional information given to participants in their respective training groups. Pairings of pictures with continua were counterbalanced across participants.

Following Maye and Gerken (2000), participants were told nothing specific about the syllables they were listening to. Diverging from Maye and Gerken, training took place over 2 days to allow for an examination of the time course of learning. Additionally, this allowed for the inclusion of more testing without disrupting the training distributions presented to the participants.

Testing During the testing phase, participants performed four tests, two focusing on perception (discrimination and identification) and two focusing on production (repetition and naming). Testing was identical for all subjects regardless of training group. Discrimination and repetition pre-tests occurred before training on each day of the experiment. At the end of each day, participants performed discrimination, repetition, categorization, and naming post-tests, though only discrimination and repetition data are presented here.

Discrimination test On each trial of the discrimination test, participants heard a pair of tokens separated by a 500 millisecond interstimulus interval and were asked to indicate whether the tokens were the same or different. Feedback was not provided between trials or at the end of the test. Stimuli 1, 3, 6, and 8 from the continua were used during the discrimination test. These tokens are presented the same number of times during training to both the unimodal and bimodal training groups. Therefore, any differences in discrimination should be due to differences in how those tokens fell in the distribution participants were exposed to only, and, critically, not to how often they heard that particular token. Presentation of pairs of tokens and the order of tokens within a pair was fully counterbalanced. Participants heard a total of 48 pairs of tokens during the discrimination test. No visual stimuli were included for this portion of the test.

The discrimination test contained three types of comparisons: same, within category, and across category. For same comparisons, participants heard one of four acoustically identical pair types: 1–1, 3–3, 6–6, or 8–8. Within-category comparisons were either Tokens 1 and 3 or Tokens 6 and 8. These comparisons fall within as defined by the unimodal and bimodal distributions. Across-category comparison contained Pairs 3–6 or 1–8. These comparisons are of tokens that fall across categories as defined by the bimodal distribution, but of tokens which fall within a single category as defined by the unimodal distribution. Critically, the test tokens were presented the same number of times in each of the training paradigms, so differences in performance cannot be driven by exposure to the specific test tokens.

Repetition test In the repetition test, participants were asked to repeat stimuli from the three continua. Participants heard a single token and were instructed to repeat the token. After participants produced a token, they pressed a button to advance to the next token. As in the discrimination test, the test tokens were Tokens 1, 3, 6, and 8 along the continuum. Four tokens of each of these points were presented. Presentation of stimuli was fully randomized. Participants were presented with a total of 48 tokens during this test. No visual stimuli were presented during this portion of the test.

Voice onset time and whole-word duration was measured for each token by two trained coders. Each coder marked burst onset, voicing onset, and end of vowel. If any amount of prevoicing was present before the burst onset, the onset of this prevoicing was also marked. Furthermore, if there were breaks between the prevoicing, and the onset of the burst, the offset of prevoicing was also marked (see Davidson, 2016, for a description of the various realizations of VOT by native English speakers). This allowed three measures to be calculated from each response each participant produced: the presence or absence of prevoicing, breaks in voicing during prevoiced tokens, and voice onset time (VOT; if positive, the duration between the burst onset or voicing onset; if negative, the duration between the onset of prevoicing and the onset of the burst). Interrater reliability was high (confidence interval for intraclass correlation coefficients [ICC] = .965 < ICC < .99).

For the sake of brevity, only data from the bimodal participants are presented in the current manuscript. Unimodal participants did not demonstrate improvement in perception or production from Day 1 to Day 2, as expected, given previous results using this paradigm. Although I describe all participants from the experiment below, only those from the bimodal training group are analyzed here.

Experiment 1

Method

Participants Eighty-nine undergraduate students (62 females) participated in Experiment 1. Participants reported no speech or hearing deficits, nor did the groups of participants differ as a function of musicianship. Seventeen participants either did not complete both days of the experiment or were not native, monolingual American English speakers, leaving a total of 72 participants for analysis. Participants were divided into two training regimens, described below, with a total of 36 participants completing each training regimen. This sample size was chosen based on previously reported effect sizes for similar perception experiments, given the lack of similar studies examining production after training in a distributional learning paradigm.[4] Below, we report data from only the 36 participants who completed the bimodal training regimen.

Procedure: Perception-only training

Participants followed the general methods outlined above for training and testing. Participants were tested in perception and production at the beginning and end of each of 2 training days. During training, participants in this condition heard a token and saw the paired picture. They then pressed a button on the button box to advance to the next token. They were not required to actively engage with the token they heard during training. Participants were presented stimuli from either the bimodal or unimodal distribution described above. Assignment of these conditions was counterbalanced across participants, so 18 participants were trained in the perception-only regimen on a unimodal distribution, and 18 participants were trained in the perception-only regimen on a bimodal distribution. As stated above, only data from participants in the bimodal training groups are analyzed here.

Procedure: Perception + production training Participants in the perception + production training regimen followed the same general methods for testing and training outlined above. The primary difference between this training regimen and the one outlined above is the task during training. As in the perception-only training, participants heard a token and saw the paired picture. However, before pressing the button to advance to the next token, participants were told to repeat the token they heard. They then pressed the button on the button box to advance to the next token. As in the perception-only training regimen, participants were presented stimuli from either a unimodal or a bimodal distribution. Assignment of these conditions was also counterbalanced. Eighteen participants were trained in the perception + production regimen on a unimodal distribution, and the remaining 18 participants were trained in the perception + production regimen on a bimodal distribution. As above, I only report data from the bimodal training groups here.

Production learning predictions If participants learn to modify their productions of the training tokens as a result of their exposure, we may expect to see participants in the bimodal groups make a bigger difference in their repetitions of endpoint tokens at the end of 2 days of training than they do at the beginning of training. Specifically, we should expect to see participants producing longer voice onset times for Token 8 than for Token 1, or more prevoicing on Token 1 than on Token 8. These differences ought to increase from pretest to posttest if participants are learning to change their productions as a result of training. Differences between perception-only and perception + production training would reflect differences in modality of training. Given previous results, we anticipate that participants in the perception + production training will demonstrate some improvement in their productions from pretest to posttest.

Production learning analysis Participants produced a total of 48 tokens in each test (three continua, four points per continuum, and four repetitions per point). Participants' productions were classified into one of four types: short-lag tokens, prevoiced tokens, mixed tokens (with substantial periods of

[4] Although the sample size used in the present study is not large, we provide a replication of the results of Experiment 1 in Experiment 2, which we believe should at least partially allay concerns regarding sample size.

both prevoicing and aspiration), and mixed tokens with a pause (with a substantial period of prevoicing, a period of silence, and a period of aspiration). Only short-lag and prevoiced tokens were used for the analyses reported here; however, Table 1 shows the proportion of each type of token.

Only the end-point tokens (Tokens 1 and 8) were compared because this is where participants are expected to make the largest differences in production. Because short-lag and prevoiced tokens are bimodally distributed, voice onset times for each token type were analyzed separately. Furthermore, because relatively few prevoiced tokens were produced, participants' voice onset times were analyzed for short-lag tokens only, and for prevoiced tokens, the proportion of tokens that were prevoiced was the dependent measure. For short-lag tokens, both raw VOT and a measure normalizing VOT for vowel duration were calculated. However, the same pattern was found across the two measures. Raw VOT is reported below. Because continuum was not a significant predictor of VOT or the proportion of tokens that were prevoiced, all continua are collapsed in the plots and analyses below.

The data were analyzed using linear mixed-effects regressions for short-lag voice onset time (Baayen, Davidson, & Bates, 2008) and logistic mixed-effects regressions for proportion of tokens that were prevoiced, implemented with R package lme4 (Bates, Maechler, Bolker, & Walker, 2014). All regressions included the maximal random-effect structure justified by the model, and the random-effect structure for each model is specified below. Significance of each predictor in the linear regressions was assessed using model comparisons.

Production learning results

Figure 2 shows the results for voice onset time of short-lag tokens for the participants in the perception-only and perception + production training groups. Examining the figure, it is clear that participants in both training groups make small differences between the two end-point tokens at pretest. This is expected, given that previous research has suggested that speakers shadow VOT (Goldinger, 1998; Nielsen, 2011). Further, participants appear to make larger differences at posttest. It also appears as though this difference is larger for the perception + production training group than the perception-only training group.

The results of the mixed-effects model support this observation. The regression included training modality, training day, token number, and their interactions as fixed effects.

Table 1 Proportion of tokens that were produced with prevoicing, short-lag voice onset time, and both prevoicing and short-lag voice onset time ("mixed" tokens) for the two training groups

Training group	Prevoiced	Mixed	Short lag
Perception only	.04	.085	.875
Perception + production	.083	.051	.866

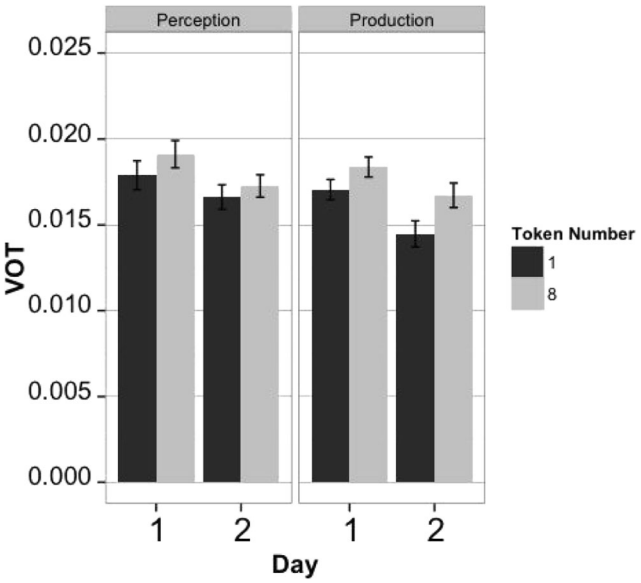

Fig. 2 Average voice onset time for short-lag tokens produced by the bimodal perception-only (left panel) and bimodal perception + production (right panel) training groups before and after training (error bars denote standard deviation)

Random effects included random slopes for training day by subject, and a random intercept for training continuum. Significance of each factor and interaction was calculated via model comparisons. First, training day was a significant predictor of model fit ($\beta = -0.0012$, $SE = 0.0007$, $t = -1.54$, $\chi^2 = 10.001$, $p = .002$), suggesting that participants changed their productions from Day 1 to Day 2. Token number was also a significant predictor of model of model fit ($\beta = 0.0024$, $SE = 0.0007$, $t = 3.378$, $\chi^2 = 68.008$, $p < .001$). However, training modality was not a significant predictor of model fit ($\chi^2 < 1$, $p = .347$).

Furthermore, the interaction between training day and token number was a significant predictor of model fit ($\beta = 0.00027$, $SE = 0.001$, $t = -1.683$, $\chi^2 = 6.178$, $p = .013$), suggesting that participants produce Tokens 1 and 8 on Day 1 rather than Day 2. The three-way interaction between training modality, training day, and token number also significantly predicted model fit, suggesting that ($\beta = 0.0028$, $SE = 0.0014$, $t = 2.015$, $\chi^2 = 4.077$, $p = .043$) participants in the perception + production training group make differences between Tokens 1 and 8 that interact with training day, but participants in the perception-only training group do not. No other interactions contributed significantly to the model fit (all χ^2s < 1, ps $> .1$)

Figure 3 shows the results for voice onset time of short-lag tokens for the participants in the perception-only and perception + production training groups. Examining this figure it is clear that participants in both training groups prevoice tokens more on Day 2 than they do on Day 1. It also appears as though participants in the perception + production training

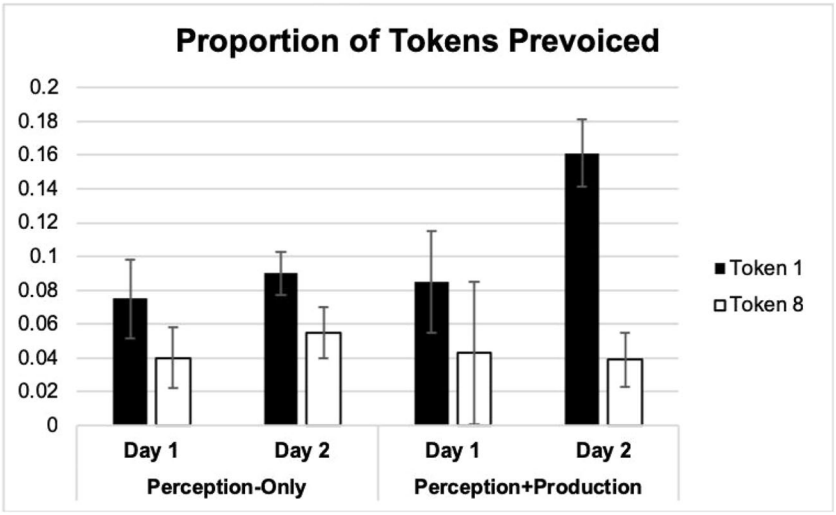

Fig. 3 Proportion of tokens that were produced with prevoicing for the bimodal perception-only (left) and bimodal perception + production training groups (right) before and after training (error bars denote standard deviation)

group produce more prevoicing on Day 2 than participants in the perception-only training group.

A logistic regression was run comparing the proportions of prevoicing across the two experiments, using the same fixed-effects and random-effect structure as the model described above for voice onset time. Because logistic mixed-effects regressions use z values, I use these estimates to determine significance. The main effect of day was significant, suggesting that participants in both training groups produce prevoicing more often on Day 2 than on Day 1 ($\beta = 1.09$, $SE = 0.41$, $z = 2.7$, $p = .007$). No other main effects or interactions were significant (z values < 1). It should be noted, however, that one should be cautious in interpreting these results, given the relatively small number of prevoiced tokens produced.

These results suggest that although participants in the bimodal perception-only training group demonstrate some changes in production, the changes are not as robust as the bimodal perception + production training group. Although perceptual training can result in production learning, more robust production learning results from training that includes production. This finding is consistent with a wide array of literature suggesting that a learner's ability to produce tokens can change with both perception and production training, but production training is more effective for eliciting changes in productions. Further, this work builds upon previous work using statistical learning, demonstrating that distributional exposure can influence production.

Perception results

Perceptual learning predictions Turning our attention to perceptual learning, I ask here whether training modality influences perceptual learning. If participants in the bimodal training group successfully learn two novel categories after training, we expect their sensitivity to across-category comparisons should significantly increase from Day 1 pretest to Day 2 posttest. However, their sensitivity to within-category comparisons should remain stable or decrease if they have a high baseline sensitivity to the contrasts. If training modality influences learning, we expect to see differences in sensitivity to across-category comparisons between the perception-only and perception + production training groups. Given previous results (e.g., Baese-Berk & Samuel, 2016), we may expect to see an attenuation of learning for the perception + production trained groups.

Perceptual learning analysis Mixed-effects models were conducted to analyze this data. Because no significant differences were found for location on the continuum to the regression (e.g., 1–8 comparisons vs. 3–6 comparisons), order of stimulus presentation (e.g., 1–8 vs. 8–1), or continuum (e.g., /da/ vs. /dæ/), these factors were not included in the models presented below, and all figures collapse over these distinctions.

Perceptual learning results

Figure 4 shows sensitivity at posttest for participants in the perception-only and perception + production training groups. Examining the figure, it appears that participants in the perception-only training group show sensitivity to the across-category comparisons, but not the within-category comparisons. It is also clear from Fig. 4 that perceptual learning is attenuated for participants in the perception + production training group.

To compare perception-only and perception + production training regimens, regressions were run that used training modality, training day, and contrast type, and their interactions as

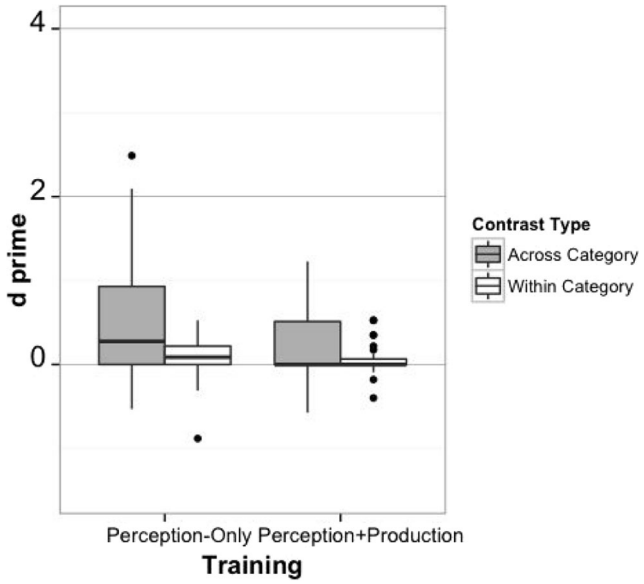

Fig. 4 Box plot of posttest performance for perception-only and perception + production training groups. Dark bars show across-category comparisons, and light bars show within-category comparisons

fixed effects. Random-effect structure was the maximal structure justified by the model (using model comparisons) and included random slopes for training day and contrast type by participant and a random intercept for continuum. Significance was determined using model comparisons (see χ^2 and p values that follow). Training day was a significant predictor of model fit (β = 0.76, SE = 0.12, t = 6.34, χ^2 = 15.01, p = .008), suggesting that overall participants performed differently on Day 1 than on Day 2. Although no other main factors were significant predictors of model fit, several interactions did emerge as significant predictors. These interactions are summarized below.

First, the interaction between training modality (i.e., perception-only or perception + production) and training day was a significant predictor (β = −0.53, SE = 0.16, t = -3.27, χ^2 = 5.36, p = .009). This suggests that the differences in performance across days are dependent on the training modality. The interaction between training day and comparison type also emerged as a significant predictor of model fit (β = −0.61, SE = 0.16, t = −3.86, χ^2 = 11.27, p < .001). This suggests that participants perform differently on across-category and within-category comparisons on Day 1 and Day 2. The three-way interaction between training modality, training day, and comparison type is also a significant predictor of model fit (β = 0.45, SE = 0.22, t = 2.03, χ^2 = 4.23, p = .039). These observations suggest that adding production to a perceptual training regimen negatively influences perceptual learning.

However, it is not the case that perceptual learning is depressed for all subjects. Examining individual data, several participants in the bimodal perception + production training group do show robust perceptual learning. A number of factors were examined as potential predictors for perceptual learning for participants in the bimodal training. Perception abilities on Day 1 during the pretest did not significantly improve the fit of the regression. It is also possible that because participants in the perception + production training group have larger demands on their attention (Baese-Berk & Samuel, 2016), and because participants do show learning in production, their perceptual learning may simply be slowed down. Perhaps with additional training time, participants in the perception + production training group would be able to learn to discriminate between the two new sound categories. This possibility is examined in Experiment 2.

Experiment 2

Method

Participants Forty-nine Northwestern University undergraduates (31 females) participated in this experiment. Thirteen participants did not complete all 3 days of training and were excluded from analysis, leaving a total of 36 participants for analysis. All participants were native monolingual English speakers and did not report speech or hearing disorders. Training groups did not differ significantly in their musical experience. All participants were paid for their participation. As in Experiment 1, each training group in Experiment 2 contained 18 participants. Participants were divided into two training groups: a bimodal perception-only training group and a bimodal perception + production training group. Because participants in the unimodal groups in Experiment 1 demonstrated no learning, we restricted training to bimodal groups for Experiment 2.

Stimuli The stimuli in Experiment 2 are identical to those in Experiment 1. Test and training stimuli are drawn from the same continua formed for Experiment 1. Because both training groups in Experiment 2 are bimodal exposure groups, there were no differences in the distributions given to participants in this study.

Procedure The procedure was identical to that in Experiment 1. All training and testing occurred in the same order as in Experiments 1 and 2. However, participants in this study trained for 3 consecutive days. The testing and training order were the same on all 3 days of the experiment. Training and testing took around 1 hour each day of the training regimen.

Production learning predictions In Experiment 1, participants in the bimodal perception-only training group did not demonstrate significant improvement in production, though there were trends toward improvement after 2 days of training. By examining repetition after 3 days of training, it is possible that

learning will emerge for perception-only training in the nontrained modality of production. If these changes do occur, it should be expected that participants in the both will make a bigger difference in their repetitions of end-point tokens at the end of 3 days of training than they do at the beginning of training. Specifically, we should expect to see participants producing longer voice onset times for Token 8 than for Token 1. Furthermore, Token 1 should be prevoiced more often than Token 8. These differences ought to increase from pretest to posttest if participants are learning to change their productions. Participants in the bimodal perception + production training should show similar patterns of learning on Day 2 as the similar group did in Experiment 1. They may also continue to improve on Day 3, showing increased differences between the two end-point tokens.

Production learning analysis As in Experiment 1, participants' productions were classified into one of four groups: short-lag tokens, prevoiced tokens, and mixed tokens (with substantial periods of prevoicing and aspiration, sometimes also including a pause). Only short-lag and prevoiced tokens were used for the analyses reported here; however, Table 2 shows the proportion of each type of token.

Participants' voice onset times for short-lag tokens were analyzed, due to the relatively small number of tokens that were prevoiced. Only the end-point tokens (Tokens 1 and 8) were compared because this is where participants are expected to make the largest differences in production. As in Experiment 1, VOT was calculated as a raw value and also as a ratio of the vowel duration. Because the two measures show similar patterns, I report raw VOT here. Additionally, the proportion of tokens that were prevoiced are also reported for Tokens 1 and 8. Because there were no significant differences across continua, all continua are collapsed in the analyses reported here.

Production learning results

Figure 5 shows the average voice onset time at Day 1 pretest and Days 2 and 3 posttest for short-lag tokens for the two bimodal training groups.

Once again, the data were analyzed using a linear mixed-effects regression that included training day, training modality, token number, their interactions, and the maximal random-effect structure justified by the model described above. The main effect of training day is significant ($\beta = 0.0038$, $SE = 0.0006$, $t = 2.556$, $\chi^2 = 35.489$, $p < .001$), as is the main effect of token ($\beta = 0.0015$, $SE = 0.0007$, $t = 2.237$, $\chi^2 = 32.581$, $p < .001$), suggesting that participants make distinctions between Tokens 1 and 8 and that their productions across training days differ. The main effect of training modality is not significant ($\beta = 0.0044$, $SE = 0.0012$, $t = 3.436$, $\chi^2 = 3.5415$, $p = .059$), though there is a numerical trend toward the participants in the perception + production training group producing longer voice onset times.

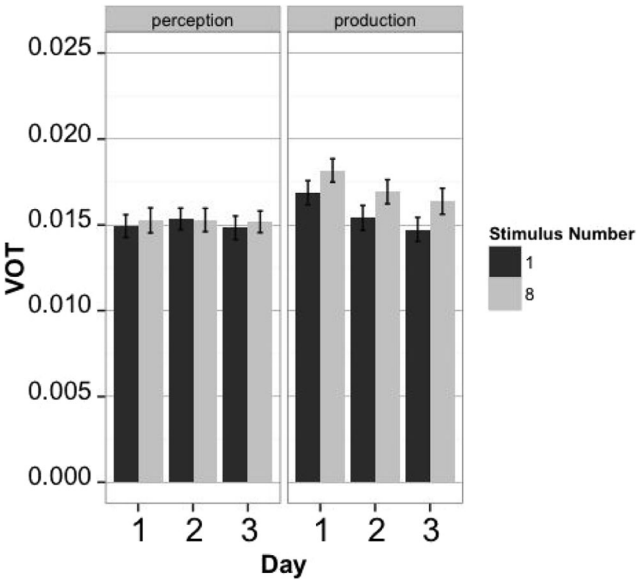

Fig. 5 Average voice onset times for the bimodal perception-only training group and the bimodal perception + production training group (error bars denote standard deviation)

In terms of interactions, only the interaction between training day and training modality emerges as a significant predictor of model fit ($\beta = -0.0045$, $SE = 0.0009$, $t = -4.726$, $\chi^2 = 30.289$, $p < .001$). All other two-way interactions and the three-way interaction were not significant predictors of model fit ($\chi^2 < 1$, $p > .1$).

As in Experiment 1, participants in both training groups make numeric differences in short-lag VOT between Tokens 1 and 8. Specifically, Token 1 is produced with a shorter VOT than Token 8. Once again, participants appear to shadow some properties of the tokens they are repeating. Although this result is slightly different than Experiment 1 in which both groups showed some differences in voice onset time between Tokens 1 and 8 after training, this lack of differences is unsurprising when examining the data. The variance in this population is large and may mask some of the very small voice onset time differences.

Figure 6 shows the proportion of tokens that were prevoiced at pretest and posttest for the two bimodal training groups.

Table 2 Proportion of tokens that were produced with prevoicing, short-lag voice onset time and both prevoicing and short-lag voice onset time ("mixed" tokens) for the two training groups

Training group	Prevoiced	Mixed	Short lag
Perception only	.05	.07	.88
Perception + production	.07	.07	.86

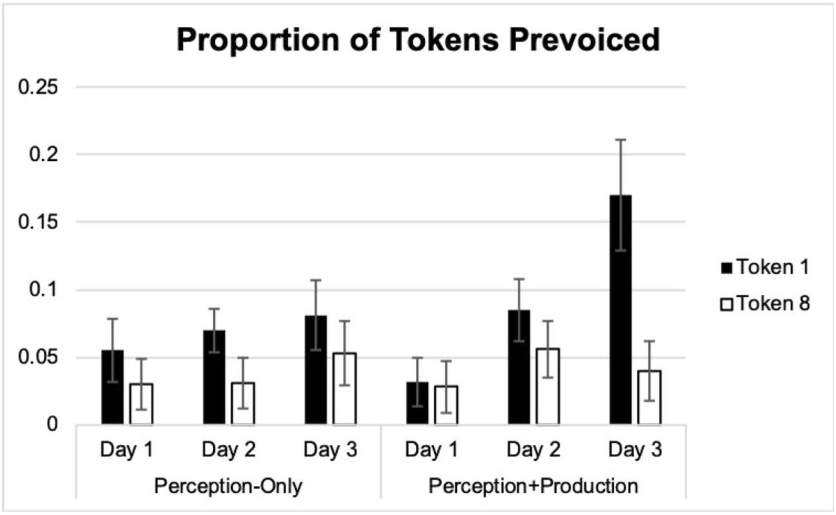

Fig. 6 Proportion of tokens that were produced with prevoicing for the bimodal training group before and after training (error bars denote standard deviation)

When examining the proportion of tokens that are prevoiced, a logistic mixed-effects model was used. Factors in the model included training modality, training day, token number, their interactions, and the maximal random-effect structure justified by the model. Training day was a significant predictor of model fit ($\beta = 2.27$, $SE = 0.34$, $z = 3.5$, $p < .001$), suggesting that participants prevoiced more often after training than they did before. Numerically, participants in both training groups do prevoice Token 1 more often than Token 8. However, token number was not a significant predictor of model fit ($z < 1$). The three-way interaction between training modality, training day, and token number ($\beta = -0.15$, $SE = 0.06$, $z = -2.17$, $p = .03$) was a significant predictor of model fit. Examining participants' performance, it is clear that participants in the bimodal perception + production training prevoiced Token 1 more often than Token 8 on Day 3 of training. Participants in the bimodal perception-only training do not make such a large distinction.

To examine this finding in more detail, follow-up regressions were run comparing Day 1 pretest with Day 2 posttest and, separately, Day 1 pretest with Day 3 posttest. No main effects or interactions emerged as significant predictors in the model examining Day 1 pretest to Day 2 posttest (zs < 1). However, in the regression that compares Day 1 pretest to the Day 3 posttest, training day is a significant predictor of model fit ($\beta = -3.8$, $SE = 1.6$, $z = -2.4$, $p = .032$), as is the three-way interaction between training modality, training day, and token number ($\beta = -3.7$, $SE = 1.8$, $z = -2.02$, $p = .043$). This supports the explanation above that the changes in productions emerge on Day 3, but not yet on Day 2. As in the case of Experiment 1, one should be cautious in interpreting these results, given the relatively small number of prevoiced tokens produced.

These results support findings in Experiment 1. Though participants in the perception-only training do demonstrate small changes in production after training, this learning is not nearly as robust as production learning after training in perception + production. Although differences were not found in short-lag voice onset time in this study, it is possible that participants were more variable in their productions in the present experiment. When examining the perception + production group independent of the perception-only group, several significant differences emerge.

Perceptual learning predictions First, participants in the perception-only training group should demonstrate robust learning after 2 days of learning, replicating the results from Experiment 1. Additionally, participants in the bimodal perception + production training group should not demonstrate perceptual learning after 2 days of training. However, on the third day, there may be a performance increase if learning in perception is simply slowed down for participants in the perception + production training group, rather than being completely disrupted. Furthermore, participants in the bimodal perception-only training group may improve their performance as a function of an increased amount of training.

Perceptual learning analysis Analyses of the discrimination data were the same as those in Experiment 1. Linear mixed-effects regressions were used to analyze the data.

Perceptual learning results

Figure 6 shows posttest scores for both the perception-only and perception + production training groups. Examining this figure, it is clear that participants in perception-only training are quite sensitive to across-category comparisons after both 2 and 3 days of training. As expected, given the results of Experiment 1, it appears as though participants in the perception + production training demonstrate less perceptual

sensitivity to the across-category contrasts after 2 days of training than participants in the perception-only training group. However, interestingly, these participants do show an improvement in sensitivity to across-category comparisons after 3 days of training, as demonstrated in the right most set of bars (Fig. 7).

To assess perceptual learning, a mixed-effects regression was performed on discrimination data from Day 1 pretest and posttests on Days 2 and 3. The regression model included the main effects of training modality, training day, comparison type, all of their interactions, and the maximal random-effect structure justified by the model with random slopes for day by contrast for participants. Contrast emerges as a significant predictor of model fit (β = −1.1659, SE = 0.2084, t = −5.594, χ^2 = 75.516, p < .001), suggesting that participants have an increased sensitivity to the across-category contrasts compared to the within-category contrasts. Neither training modality nor day significantly improves model fit (both χ^2 < 2, p > .10).

Examining the interactions included in the model, the three-way interaction between training day, training modality, and contrast type was not significant (χ^2 < 1, p > .10). However, the two-way interaction between training day and modality was significant (β = .5538, SE = 0.2908, t = 1.904, χ^2 = 5.9372, p = .015), as is the interaction between training modality and contrast type (β = 0.4857, SE = 0.2908, t = 1.67, χ^2 = 4.4422, p = .035). The interaction between training day and contrast type is not significant, though there is a numerical trend toward across-category comparisons being more distinct on Day 2 than on Day 1 (β = −.3149, SE = 0.2948, t = -1.069, χ^2 = 3.4823, p = .06).

In Experiments 1 and 2, participants show improvement that is largely tied to the modality of training. Participants trained in perception + production demonstrate substantial improvement in repetition from pretest to posttests, but only begin to show gains in perceptual learning after 3 days of training. Participants trained in perception-only demonstrate improvement in discrimination, but do not demonstrate learning in production. After 3 days of training, they do show some improvement in production, but these advances are relatively limited compared to participants in the perception + production training group.

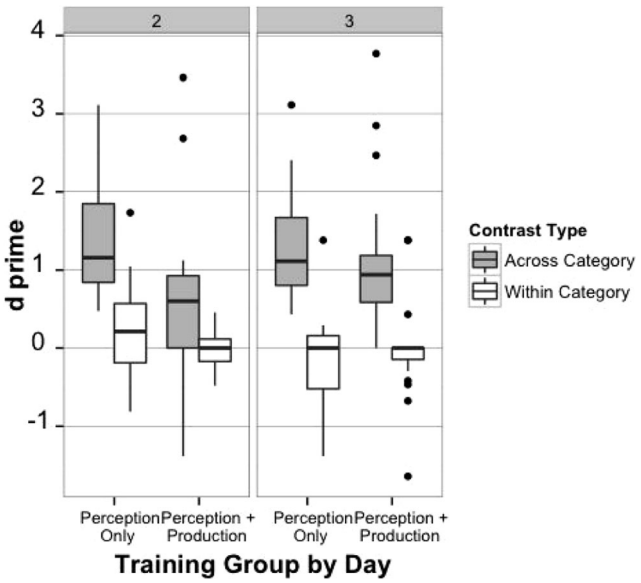

Fig. 7 Participants' d' scores after 2 and 3 days of training

At first blush, this is not surprising. Training focused on a particular modality should show substantial improvement within that modality. However, one aspect of this finding is rather puzzling. How do participants in the perception + production training learn in production when they are unable to perceive differences between the training tokens? This is particularly curious because the training task was a repetition task, which requires the learner to perceive the token they are trying to produce. Although the results of Experiment 2 suggest that perceptual learning is not entirely disrupted, the finding of reduced learning after 2 days of training in perception merits further review. In the next section, I investigate individual variability in perceptual learning and whether perceptual learning correlates with production learning for the training groups in Experiment 1. I then examine several possible factors underlying individual variability in perceptual learning.

Individual differences in perceptual learning Figure 8 shows individual performance for the bimodal training groups from Experiments 1 and 2. Individual performance on the across-category comparisons during the Day 2 posttest are plotted here; participants are ordered by their final performance on the discrimination task. A few interesting observations can be made. First, there is substantial individual variation across participants in both training groups. However, focusing on the perception + production training group in the top panels of Fig. 8, it is clear that 11 of the 18 participants are performing at chance levels on the discrimination task, even after 2 days of learning.

This pattern was also seen in the individual performance for participants in Experiment 2. The individual data are plotted in the bottom panels of Fig. 8. Participants are ordered by their performance on the discrimination task on Day 2. Interestingly, some participants in the perception + production training group who do not learn after 2 days of training do demonstrate improvement from Day 2 to Day 3. However, a small number of participants still demonstrate very poor performance (i.e., chance performance on the discrimination task) even after 3 full days of training. This suggests that the source of disruption may not simply be a delay in learning, which could be alleviated by increased or prolonged exposure. I examine some possible options for this disruption below.

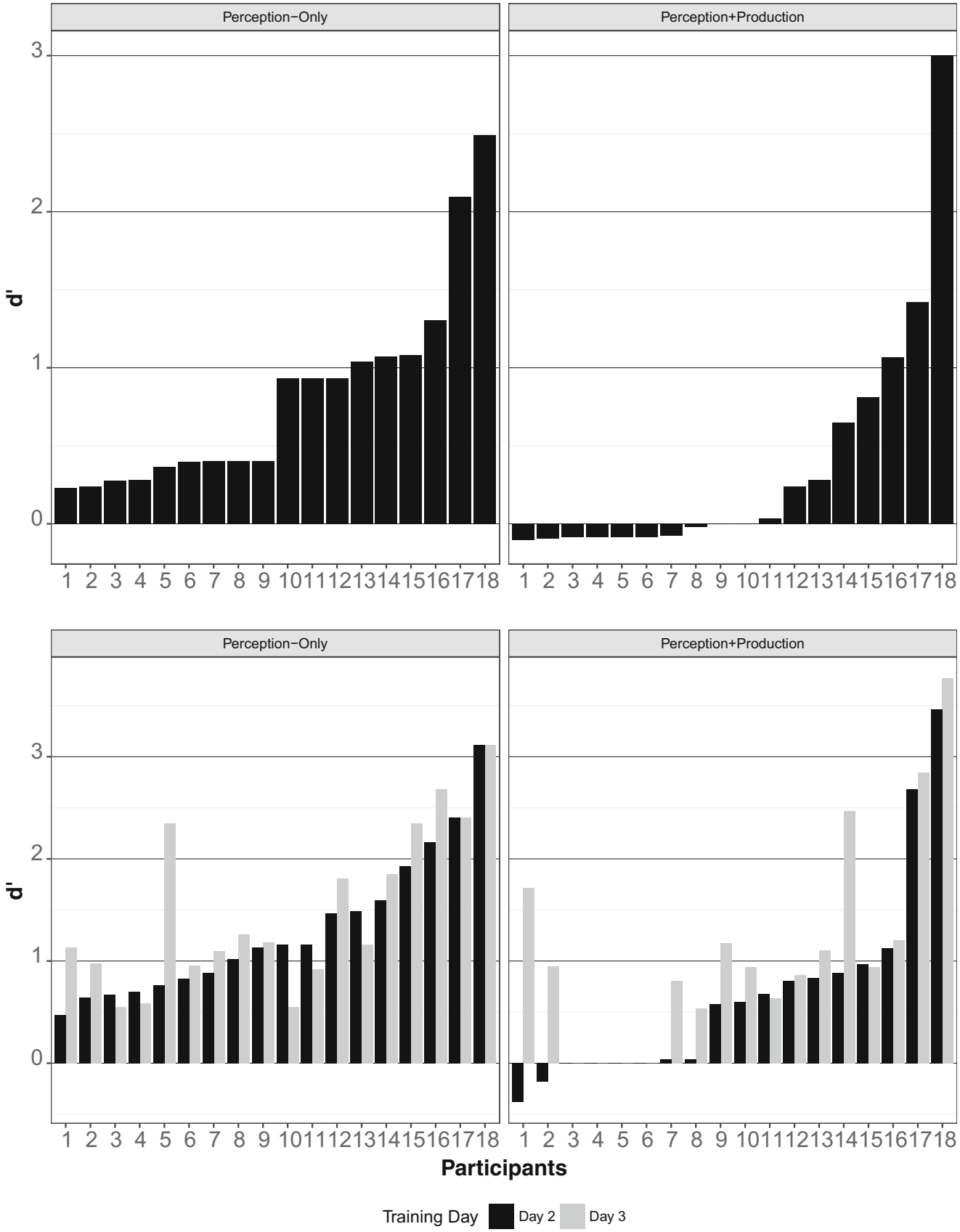

Fig. 8 Individual performance for the bimodal training groups in Experiment 1 (perception only: top left panel; perception + production: top right panel) and Experiment 2 (perception only: bottom left panel; perception + production: bottom right panel) on the discrimination posttest

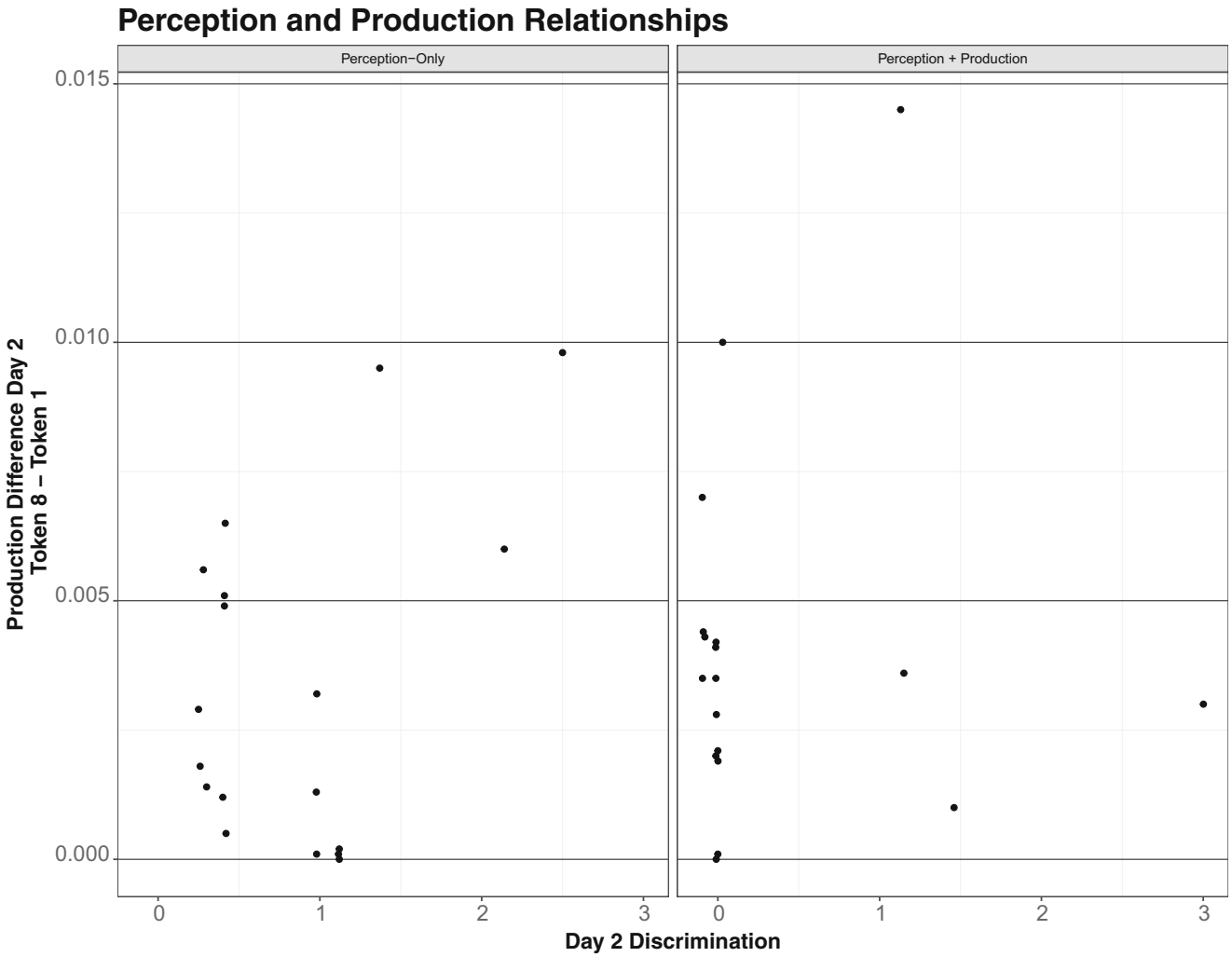

Fig. 9 Day 2 discrimination (*d'*) and the amount of difference in voice onset time between Tokens 1 and 8 on Day 2. Perception-only training is shown in the left panel, and perception + production training is shown in the right panel. Discrimination is shown in *d'*, and the production data are shown in VOT (seconds)

One may wonder whether participants who fail to learn in perception also fail to learn in production. It is possible that the gains in production learning for the perception + production training group are driven by the few participants who demonstrate perceptual learning. In order to ask this question, I examined correlations in perception and production for both training groups in Experiment 1.

To examine the correlation in learning across modalities, model comparisons were performed for models that included Day 2 discrimination performance as a predictor of the difference participants make between tokens in production on Day 2. The model that included Day 2 discrimination was a significantly better fit than the model that did not include that comparison ($\chi^2 = 11.6, p < .03$). This suggests that performance in the two modalities is related for participants in the perception-only training group. Figure 9 shows Day 2 discrimination performance and the amount of difference made between Tokens 1 and 8 in production.

A similar comparison for the perception + production training group reveals a different pattern. The model that included Day 2 discrimination was not a significantly better fit than the model that did not include that factor ($\chi^2 = 8.1, p = .09$). This suggests that for the perception + production participants, performance in production is not related to performance in perception. The right panel of Fig. 9 shows Day 2 discrimination and Day 2 repetition performance for the perception + production training group. Of particular interest are the 11 subjects who do not learn in perception (clustered around the zero point on the horizontal axis). Several of these participants demonstrate production differences on Day 2, suggesting that learning in production is not tied to learning in production.

Although it is clear that performance on perception and production tasks is not closely tied for the participants in the perception + production training, a number of other factors could influence the disruption of perceptual learning. Several of these possibilities were examined using model

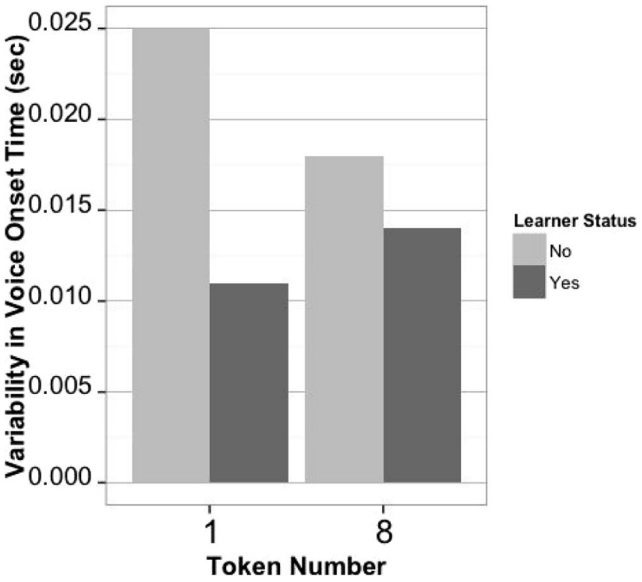

Fig. 10 Variability in voice onset time for the perception + production training group in Experiment 1, divided by learners and nonlearners

comparisons. Day 1 pretest performance on the discrimination task does not significantly predict model fit for the perception + production training group ($\chi^2 < 1$).

One intuitive explanation is that participants who were "bad" at the repetition task were also "bad" learners. That is, participants who gave themselves productions that deviated from the distributions given to participants during training may have caused a disruption in their own perceptual learning. Taken in its most basic form, this does not appear to be true. As demonstrated above, examining only the perception + production training group, Day 2 posttest production performance does not significantly improve model fit. Baseline production abilities also do not significantly improve model fit ($\chi^2 < 1$), nor does average performance on either of the two end-point tokens ($\chi^2 < 1$).

However, it is possible that average production abilities do not accurately characterize what participants produce during test and training. Substantial previous work has suggested that variability plays a critically important role in learning. In order to examine the role of production variability during learning, I measured all productions of Tokens 1 and 8 during training and test.[5] I then calculated the variability for each token. For visualization purposes, participants in the bimodal training group into two groups: learners and nonlearners. Participants were grouped as a function of whether their *d′* values at posttest for the across-category comparison were above chance (learners) or at chance (nonlearners). Figure 10 shows the variability in voice onset time for Tokens 1 and 8 for learners and nonlearners. This figure shows that participants who do not learn in perception are more variable than participants who do learn in perception, specifically on Token 1. Recall that Token 1 is at the prevoiced end of the continuum. It appears that participants are more variable on the end of the continuum that is less frequent in English and is likely more novel to participants in the present study.

The results of a linear mixed model support this observation. A mixed model was run with the variability in voice onset time as the dependent variable, and learner status (i.e., learner vs. nonlearner) and token (1 vs. 8) and their interaction as fixed factors, and the maximal random-effect structure justified by the model. A model comparison demonstrated inclusion of learner status significantly improves model fit (χ^2 = 4.517, p = .033), but inclusion of token and the interaction of token and learner status do not improve model fit (both $\chi^2 <$ 1).

General discussion

The results of these studies replicate previous findings that participants trained in both perception and production demonstrate disrupted perceptual learning as compared with participants trained in perception alone. Specifically, participants trained in perception alone demonstrate robust perceptual learning. Participants trained in both perception and production demonstrate robust learning in production, but less learning in perception. Greater improvement in the modality that was the focus of training is not a particularly surprising finding. However, it is rather surprising that participants in the bimodal perception + production training learn to produce tokens more accurately after training, even though they do not show evidence of perceptual discrimination between these two categories. An additional key finding, further differentiating this work from Baese-Berk and Samuel (2016) is that a third day of training partially alleviates the disruption to perceptual learning after training emphasizing production. However, substantial individual differences in perceptual learning remain, even after a third day of training. Therefore, a key question about this data is *why* perceptual learning is disrupted when participants produce tokens. An investigation of the results reveals several unlikely sources for the disruption, as well as some possible avenues for future research.

Below, I outline the implications of this work for our understanding of variability during training and how perception and production may be susceptible to different types of learning. I also discuss the implications of these results for our understanding of the relationship between perception and production, specifically during learning, and outline a proposed framework for understanding how the two modalities interact with one another.

[5] Here, I only present data with short-lag VOT; that is, the voice onset time of prevoiced tokens is not included here because it would highly skew the results (i.e., most prevoiced tokens have a very long period of prevoicing). Further, it is not the case that the participants who do not learn in perception also produce the most prevoiced tokens, suggesting prevoicing alone cannot account for our results.

Variability in production and learning

The results in the present study revealed no significant correlation between perception and production abilities for the participants trained in perception + production, suggesting that the lack of perceptual learning is not driven only by participants who also do not demonstrate learning in production. This is supported by analyses suggesting that neither participants' baseline production abilities nor their production abilities at posttest predict their performance on the perception tasks. It is also not the case that participants baseline perception abilities are predictive of performance in perception after training. However, an examination of variability, especially variability on the "new" token seems to at least partially predict performance on the perception task. That is, increased variability in production correlates with a disruption in perceptual learning.

These results suggest that variability influences performance in a different way than absolute accuracy. Before beginning this discussion, it is important to note that variability can occur in a number of different forms: variability in learner performance, acoustic variability in productions from speakers, acoustic variability in the input to listeners, semantic variability, and speaker variability. Although it is clear that each of these types of variability have properties that are quite different from one another, substantial previous work has treated variability as a monolithic construct. Below, I describe some previous work on variability, particularly with regard to nonnative speech learning, and address how the results of the present study fit into this prior work.

Some previous work has suggested that nonnative speech can be more variable than native speech (e.g., Baese-Berk & Morrill, 2015; Wade, Jongman, & Sereno, 2007); however, other work suggests that, under some circumstances, nonnative speech is *less* variable than native speech (e.g., Morrill, Baese-Berk, & Bradlow, 2016; Vaughn, Baese-Berk, & Idemaru, 2018). This is, perhaps, unsurprising because native speakers produce substantial amounts of variability. Therefore, it may be more appropriate to reframe the notion of "correct" production. Instead, nonnative speakers must learn how to appropriately deploy variability in their productions and interpret variability in their production.

Taking this reframing, these results have interesting implications for how variability affects learning. If a learner must be able to acquire appropriate variability, but variability correlates in this study with a disruption of perceptual learning, how can we reconcile this conflict? In fact, this study is not the first to recognize such a conflict. Substantial prior research has demonstrated two conflicting consequences of variability during training—some in which it hinders learning and some in which it helps learning (e.g., Barcroft & Sommers, 2005; Sommers & Barcroft, 2007). One critical question is under what circumstances variability may have a positive impact and under what circumstances it can be disruptive.

In general, variability is thought to be one of the primary challenges of speech perception. The listener must determine what variability in the acoustic stream is meaningful for phonetic contrasts and what variability is not meaningful for phonetic contrasts.[6] However, it is also important to understand the range of variability that a particular phonetic feature may map on to. Substantial previous work has suggested that variability in both words and voices can benefit lexical learning and phonetic learning for nonnative speakers (Bradlow et al., 1997; Iverson et al., 2005; Sommers & Barcroft, 2007). Further, it is clear that exposure to multiple speakers can help listeners better understand a novel speaker from either a familiar (Bradlow & Bent, 2008; Sidaras, Alexander, & Nygaard, 2009) or unfamiliar (Baese-Berk, Bradlow, & Wright, 2013) background. However, additional work by Barcroft and Sommers (2005) has demonstrated that some types of variability can hinder the learning of novel words. Specifically, variability in the semantic representation learners are exposed to can disrupt learning, whereas variability in the form tends to enhance it.

In the present study, the variability being considered is always in the form, but it is still possible that it differs from variability in studies discussed above in important ways. Specifically, in previous studies, the experimenter has controlled the amount of variability that a listener is exposed to. This is true even in the distributions provided to learners in the present study. However, what is not controlled in the present study is the variability learners are exposed to in their own productions. It is possible that experimenter-given variability is structured in a particular way such that the variability itself is more stable. However, when listeners are exposed to their own voices, is it possible that this variability differs in its structure. That is, the variability could be even less predictable for the learner than the variability they are exposed to in other studies, where it is more carefully controlled by the experimenter. Further, it is possible that listeners weigh variability in their own productions differently than variability they are exposed to in other productions. When examining adaptation to an unfamiliar speaker or production, Sumner (2011) suggests that the type of variability is likely to affect how learning and adaptation proceed, which is consistent with the results of the present study. Previous work has also suggested that variability may facilitate learning when that variability is tied to indexical features (e.g., Rost & McMurray, 2010); however, when variability is within a contrastive acoustic dimension it can disrupt learning (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Holt & Lotto, 2006; Lim & Holt, 2011).

[6] Although variability is not informative for the phonetic contrasts here, it is informative for other aspects of processing the speech signal, such as indexing a particular speaker's voice.

Therefore, perhaps the variability we see in the present study (within a contrastive acoustic dimension) could be a source of disrupting perceptual learning precisely because it is within a contrastive acoustic dimension.

It is also possible that the tasks themselves could directly impact variability. Previous work has demonstrated that dual tasks increase variability in speech motor performance (Dromey & Benson, 2003), and the production + perception condition in the present study is an example of a dual task. However, it should be noted that in the case of variability as it is investigated here, all participants are completing the same tasks; therefore, the explanation that dual tasks increase variability cannot entirely account for our results. That is, participants in the perception + production training were differently variable from one another, even though they were completing the same tasks. That said, it is possible that dual tasks have a greater impact on some learners than others, with regard to variability in production.

Different types of learning in perception and production

As discussed above, understanding variability is critically important for understanding category learning. But one must ask whether participants in this experiment are learning categories at all, or whether different types of learning may be emerging in different training groups. The type of discrimination tested here requires learners to develop category representations, as learning was defined as an increase in across-category discrimination from pretest to posttest, but critically not within-category learning. Of course, it is possible that learners could improve at both within-category and across-category distinctions; however, this would suggest that their overall discrimination was improving and would not suggest that they were forming two novel categories in perception. That is, a lack of learning in perception could imply a lack of learning, or it could imply simply a different type of learning. Rather than acquiring a novel category, the learner may be more proficient at fine-grained discrimination. Obviously, this skill is less useful for phonological category learning, which requires a listener to be able to generalize over irrelevant variability to acquire a novel category. However, to imply that this is not learning may be an inappropriate interpretation of the results.

Whether phonologically categorical learning is occurring at all for participants in the present study is, in fact, unclear. Some previous work has directly addressed the issue of what is required during repetition and whether phonological categories are reflected in imitation, especially, for a second language. For example, Hao and de Jong (2016) examine phonological mediation during imitation, and demonstrate that second-language (L2) speakers show little evidence for phonological encoding during this task. This suggests that perhaps imitation or repetition in L2 does not directly require phonological (i.e., categorical) information. Therefore, a learner could improve at repetition in two ways. One is to acquire a novel category and to be able to more accurately select an exemplar for production from that category. However, a learner could also improve at repetition simply by more accurately matching their production to the token they hear. This would not require formation or access of a category. Naming, on the other hand, requires learners to access categorical representations, and improvement of the second type above would not allow for improvement in naming performance.

Similarly, in perception, it is possible that learners are not actually acquiring phonological (or even linguistic) categories, even in the bimodal case. That is, it is possible that a learner is simply acquiring the fact that the distribution of sounds is bimodal, but not necessarily that sounds in each of the two modes are from distinct categories. In fact, this is all that is required for learning to occur in some models (see, e.g., Kronrod, Coppess, & Feldman, 2016). Therefore, it is necessary to further probe linguistic knowledge, especially phonological category knowledge, in order to determine what, exactly, participants are learning. This suggests that a wide range of skills should be tested in order to better understand whether learners truly are acquiring novel categories.[7]

It is also possible that the results of the present study are, in part, driven by the specific contrast presented to participants. Previous work has suggested that some American English speakers produce prevoicing for word-initial voiced stops, even if this prevoicing is not contrastive (e.g., Davidson, 2016). If this feature is used more by some speakers than by others, it is possible that it is indexically informative to many listeners. Therefore, differentiating between prevoiced and short-lag stops as being phonologically informative may be a challenge for listeners who are exposed to many speakers who use prevoiced stops. Similarly, if a speaker is more likely to produce prevoicing themselves, it is possible that this may influence their ability to learn to perceive and produce prevoiced tokens as distinct from short-lag tokens. That is, a learner who naturally produces many prevoiced stops may be able to use preexisting motor plans to produce prevoicing more effectively than a learner who naturally produces very few prevoiced stops. Alternately, a learner with more experience producing prevoiced stops may have a more difficult time controlling prevoicing and differentiating it from short-lag stops, as they may not reliably do so during typical production. This is unlikely to be the sole driver of our effects, as the number of participants who prevoiced tokens at pretest did not differ across our training groups.

[7] Some previous work has suggested that in order for learning to truly be interpreted as "category learning," learners must demonstrate generalizability to novel talkers or novel contexts for the contrast.

Of course, it is possible that each modality may be differently susceptible to category learning and the impact of variability on category learning. For example, in order to accurately perceive speech, a listener must be able to determine the target over substantial variability ranging from tokens from unfamiliar speakers to tokens from the same speaker produced under different circumstances. Perception also relies on what the learner is exposed to—that is, the learner cannot control what he or she hears in perception. Production, on the other hand, allows the learner relative autonomy. In normal conversation, speakers can choose, in most circumstances, what they want to say and when they want to say it. Therefore, one could imagine a circumstance in which perception requires flexibility and is quite susceptible to learning, whereas production is more inflexible and less susceptible to change.

In fact, previous proposals about perceptual learning have directly addressed the fact that the perceptual system must be flexible; but this flexibility comes with a cost. While discussing perceptual learning, Samuel (2011) argues that "change is necessary; change is bad." He discusses both the requirement for and the costs of substantial change in the perceptual system. In the case of perceptual learning, change is something that must occur in order for the perceptual system to appropriately interpret the input. However, one could imagine extending this argument to production and stating that change is bad, but only sometimes necessary. That is, even in cases where the learner slightly mispronounces some phoneme, it would rarely result in serious misunderstandings. Further, if the production system were as flexible as the perceptual system, one might expect that the changes in production would render a speaker unrecognizable from one production to the next. The production system should not be as flexible as the perceptual system, so it is possible that the two systems are differently susceptible to change, broadly speaking. Further, it may be the case that when changes in production occur, changes in perception are relatively attenuated, in order to maintain stability in one section of the system.

It is also important to note that the relationship between perception and production, and the relative plasticity of each modality, may differ depending on the target of learning. For example, Leach and Samuel (2007) demonstrate that production improves some aspects of perceptual learning, but hinders other aspects. Learners better acquired formal properties of the word, termed "lexical configuration." However, "lexical engagement," or how the target word interacts with other words in the lexicon, was hindered by production during training. These results suggest that consideration of multiple targets of learning are necessary to fully understand perception and production learning and how they operate together.

Further, Thorin, Sadakata, Desain, and McQueen (2018) examined production of related and unrelated tokens during learning. Their results demonstrate no differences between the two groups as a function of training group (i.e., production of related or unrelated tokens). Both groups improve equally well in perception and in production. This result contrasts, in some ways, with the current findings. However, the Thorin et al. paradigm provides some control for cognitive load, a feature not controlled in the current study, which is a known factor to influence this type of learning (Baese-Berk & Samuel, 2016). Below, I propose an account that could elucidate the broad spectrum of results examining perception and production.

Relationships between perception and production during learning

Although it is possible that a number of factors may modulate the relationship between perception and production and how amenable each of these categories is to learning, the fact remains that in order to successfully use a language, an individual must be able to perceive and produce sounds from categories. Therefore, it is important to examine how these results speak to models of second-language learning, as well as to the relationship of perceptual and action modalities more broadly speaking.

The two most prominent theories of acquisition of nonnative speech sounds (SLM: Flege, 1995; PAM-L2: Best & Tyler, 2007) cannot easily account for these data. Both models assume that many of the difficulties in perception and production are shared and that learning in one modality should mirror learning in the other. The present study demonstrates that perception and production learning are strongly correlated after training in perception alone; but after training in production, there is no relationship between the two modalities, even after 3 full days of training. Further, the current studies demonstrate several dissociations between the two modalities, which runs counter to the predictions of these two theories.

A further prediction of the speech learning model is that perceptual learning should lead production abilities, and that is not the case in the present data. Even examining the relationship between perception and production in perception training, perceptual learning does not always precede production learning. This is in line with many, many previous studies that show great individual differences in terms of which modality is learned first.

One issue not addressed in the present study that is a major focus of both SLM and PAM-L2 is the relationship between the first and second language. It is possible that the relationship between perception and production may shift as a function of the relationship between the learner's native language and the target language. Some contrasts may be more salient in perception and some may be easier to articulate. Further, whether similar contrasts exist in the L1 and L2 may influence how easily the sounds are learned and may modulate the relationship between the two modalities. It is also possible that the relationship between the two modalities may shift over time;

however, it is important to note that Baese-Berk and Samuel (2016) demonstrate a significant disruption to perceptual learning, even for learners with substantial experience with the nonnative contrast, even though this disruption is substantially smaller than it is for naive learners.

In addition to considering the implications of these results for models of second-language learning, it is also important to consider the implications of these results for our understanding of speech perception and production more broadly. Previous work has outlined three primary perspectives on perception: direct realism, motor theory, and a general auditory account (Diehl et al., 2004). Direct realism and motor theory make clear predictions about how perception and production ought to be correlated, since both posit that listeners perceive gestures (or intended gestures). A general auditory account (Diehl & Kluender, 1989), however, accounts for relationships between perception and production using different mechanisms. Diehl et al. (2004) summarize these approaches as positing that perception follows production and production follows perception. That is, the two systems work in concert to shape each other. For example, the need for auditory distinctiveness among sounds constrains the production system (Kingston & Diehl, 1994, 1995; Liljencrants & Lindblom, 1972). Similarly, listeners integrate auditory and visual information during perception (McGurk & MacDonald, 1976). The results presented here are broadly consistent with a general auditory account of perception, which can provide an explanatory framework for complex relationships between the two modalities.

Time course of learning in perception and production

The results of the present study have important implications for our understanding of the time course of learning in perception and production. With an additional day of training, the disruption in learning between the two modalities was alleviated for many participants, suggesting an increase of training may aid in connecting learning in each modality. This result echoes a finding in Baese-Berk and Samuel (2016), who demonstrated that perceptual learning was disrupted to a smaller degree for individuals who already had some familiarity with Basque, the target language. That is, learners who had more exposure learned more in perception than learners without such exposure. Outside the laboratory, in more naturalistic classroom settings, similar findings have been demonstrated. Nagle (2018) shows that production learning is delayed as compared with perception learning for some aspects of an L2 contrast, but not for others. Further, he demonstrated substantial individual differences in the strength of correlation between performance in perception and production.

Taken together, these sets of findings suggest that the relationship between the two modalities is likely not static, shifting over time as a function of exposure and learning in each modality. If this is the case, it is likely that later learners may show a different pattern of perception–production interactions than novel learners. Some preliminary evidence for this prediction comes from Thorin et al. (2018), who demonstrate no correlation between perception and production learning for Dutch learners of British English, who have substantially more experience with the target contrast than the naïve learners in the current study, or even, presumably, than the late learners of Basque in Baese-Berk and Samuel (2016). A combination of approaches, examining both short and long time scales, are necessary to fully understand the dynamic nature of this relationship over time.

Implications for second-language learning outside the laboratory

It is also clear that this work has serious implications for how nonnative languages are taught. Many modern models of teaching focus primarily on rapidly achieving communicative competency (N. C. Ellis, 2003, 2009; Hymes, 1972; Nunan, 2002). In order to achieve this, many instructors require students to produce tokens very early on in their learning experience. For example, early repetition of words is emphasized in many second-language classrooms and is advocated by many researchers and in many teacher-training programs (e.g., Brown, 2015). However, the results in the present study suggest that early production, especially for naïve learners, may be harmful for a learner's ability to eventually perceive the differences between some contrasts.

Some research in second-language teaching from the late 1980s advocated for a different approach to L2 teaching. For example, Krashen (1985, 1989) claims that a "silent period" naturally occurs for adult learners, similar to the experience of many immigrant children experience upon arrival in a new environment that does not share their native language (Krashen, 1981) However, this hypothesis fell out of favor, and a focus on input shifted to a focus on output. R. Ellis and Shintani (2013) note this lack of attention to input in second-language acquisition pedagogy. They argue that the focus on communication and output has resulted in insufficient focus on input in professional development and books on pedagogy. In recent years, there has been a shift back toward a focus on input (e.g., McDonough, Shaw, & Masuhara, 2013; Polio, 2007), though a focus on production early in the learning process remains. It is possible that depending on the goals of learning, delaying production until a learner has more experience with perception of the language may aid in later perceptual development. It is worth noting that difficulty in perceiving and producing many notoriously challenging contrasts persists even at relatively advanced stages of L2 acquisition.

All this said, it should be noted that these experiments are laboratory based and were conducted in a controlled environment, which differs significantly from most real-world

language acquisition, and the speculations above should be taken as that, rather than statements of fact. Whether taking place in a classroom or in a naturalistic setting, second-language acquisition includes many variables that were not manipulated here. Therefore, in a more ecologically valid environment, the relationship between the two modalities may different than the relationship shown in the present study (though, see Nagle, 2018, for naturalistic evidence that the relationship between perception and production is complex during naturalistic learning). More work should be done to examine how perception and production influence each other in more naturalistic language settings.

Shared resources account for perception and production learning

Although the bulk of the discussion thus far has focused on the relationship of speech perception and production during learning, these results also have implications for the relationship between perception and action more broadly speaking. In particular, the disruption of perceptual learning after production training is informative about this relationship. If perception and production were to be entirely dissociated, such a disruption would be quite surprising. However, if perception and production were identical, a disruption would similarly be unlikely. Given that perception and action must be separate in some ways but linked in others, it is important to begin to investigate how this relationship may manifest. I return to the specific case study of speech perception and production here to provide a potential account for the dissociation and transfer seen in the present study.

To help understand the observation of both dissociations and transfer between learning in these two modalities, I develop an account that appeals to shared resources across these modalities. Ferreira and Pashler (2002) propose an account relying on a central bottleneck theory to explain interference during word production. They suggest that if two tasks share processing resources, and a stage of one task requires central processing resources, the second task will not also be able to use those resources until the first task has completed its process. Under this hypothesis, if production and perception share resources, trying to perform both perception and production simultaneously or in quick succession may result in a bottleneck of processing resources, slowing down or hindering the task. Below, I outline an account for the relationship between perception and production that appeals to a resource-sharing hypothesis.

Under this account, the representations for perception and production at the phonetic level are separated. Perceptual learning that is driven by perception training recruits cognitive processing resources. Once new representations have been established in perception, this learning can partially transfer to help form new representations in production. That is, though resources are shared across the two modalities, because only one modality is emphasized during training, learning progresses in that modality, and once learning is sufficient, it can transfer to the other modality.

Learning during perception + production training requires resources to be split between the two modalities; essentially, it is a type of dual task. That is, during the perception + production training here, participants are asked to both perceive and produce tokens on every trial. Further, production itself is a more costly task in terms of processing resources, as compared with perception alone. In the dual task in the current study, learning in production occurs by establishing new representations in production. However, because resources are divided between perception and production, and because production is very resource demanding compared with perception in this case, the formation of perceptual representations is slower than after training in perception alone. Further, it is also possible that production learning is a slower process, a result suggested by the large improvements on Day 3 of training in Experiment 2 for the perception + production training group. If this is the case, it is possible that learning is not sufficiently strong after 2 days of training to allow the use of resources for perceptual learning.

This account also allows us to understand both the dissociation and transfer between learning in the two modalities. Because the representations in the two modalities are formed by different processes within each modality during learning, there is a dissociation between performance in each modality for the perception + production training group. In contrast, during the perceptual training task, distinct production processes are not recruited during learning. Because perception-only training is not as resource demanding as production, learning in one modality can transfer to the opposite modality. Resources that would otherwise have to be split between perception and production can be used for perceptual learning and transferring that learning to production, accounting for transfer between the two modalities after perception-only training. Unfortunately, in the present study, participants did not complete any cognitive tests, and thus it is impossible to more directly asses the resource-sharing account, given the present data. However, some recent data provide evidence for this hypothesis, demonstrating that perceptual learning is disrupted in other types of dual tasks that do not involve direct production of the target (Baese-Berk & Samuel, 2016). This hypothesis also predicts that increasing the cognitive load during either perception-only training or perception + production training should increase the disruption to perceptual learning. Future studies should more directly assess cognitive measures (e.g., executive function, attention) to provide support or modification to this account.

Of course, myriad factors will influence how perception and production interact both during learning and during processing of learned language, including previous knowledge, attention,

timing, and a variety of other cognitive and noncognitive factors. For example, autonomic responses are heightened during speech tasks (see, e.g., Arnold, MacPherson, & Smith, 2014; Francis, MacPherson, Chandrasekaran, & Alvar, 2016). However, it is an open question whether or not this increase in autonomic response might play a role in the disruption of learning seen in the present study. An examination of the influence of these responses, and their interactions with the cognitive factors described above could be examined in future research. One could also imagine that this relationship could be manipulated during training by manipulating factors including to which features (or tasks) the listener's attention is being drawn (see, e.g., Pederson & Guion-Anderson, 2010). Some of the factors that may affect learning may be directly involved in the shared processing account here (e.g., attention, working memory), whereas others may influence learning through other routes.

As a final note, it is important that future accounts of the relationship between perception and production during learning be able to account for shifts in this relationship over time. That is, while the present study demonstrates a disruption to perceptual learning after training in production, it is not the case that one should expect the relationship between the two modalities to remain antagonistic as learning progresses. Instead, these results, and others, should be taken as evidence that the relationship between the two modalities can take a particular form under a particular circumstance. Shifting those circumstances should, almost certainly, shift the relationship, and future research should attempt to address how this relationship shifts as a function of the circumstances of learning.

Conclusion

The experiments in this study were designed to examine the relationship of perception and production during learning. The primary objectives of this study were to examine the role of training modality in learning and to examine whether learning in one modality is related to learning in the other modality. The data presented here demonstrate that learning can, but does not always, transfer to between modalities. Further, learning in production is not dependent on learning in perception occurring first. These results suggest that theories explaining how perception and production are related must be constrained in a variety of ways. It is likely that the representations in the two modalities are formed separately, but that some processes allow for transfer of information and learning between the two modalities. This transfer process may be nonobligatory and is, at the very least, not automatic for all cases of learning. Future studies of the relationship between perception and production should examine why the two modalities interact in these ways and how the relationship between these two modalities evolves over time such that learners shift from novice listeners to masters of a nonnative language and the language's phonological contrasts.

Acknowledgements This work was supported by National Science Foundation Grants BCS-0951943 and BCS-1734166. I would like to thank Ann Bradlow, Matthew Goldrick, and Arthur Samuel for their comments on previous versions of this work.

Open practices statement The data reported here are available, but none of the experiments were preregistered.

References

Arnold, H. S., MacPherson, M. K., & Smith, A. (2014). Autonomic correlates of speech versus nonspeech tasks in children and adults. *Journal of Speech, Language, and Hearing Research, 57*(4), 1296–1307.

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

Babel, M. (2011). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 1–13. https://doi.org/10.1016/j.wocn.2011.09.001

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*(1), 1–28. https://doi.org/10.1515/lp-2014-0006

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

Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *Journal of the Acoustical Society of America*, *138*(3), EL223–EL228. https://doi.org/10.1121/1.4929622

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(3), 387–414.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 [Computer software]. Retrieved from https://rdrr.io/cran/lme4/

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Best, C. T., McRoberts, G. W., & LaFleur, R. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development, 18*(3), 339–350. https://doi.org/10.1016/0163-6383(95)90022-5

Best, C. T., McRoberts, G. W., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345–360.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. S. Bohn (Ed.), *Language experience in second language speech learning in honor of James Emil Flege* (pp. 13–34). Amsterdam, The Netherlands: John Benjamins.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer software]. Retrieved from http://www.fon.hum.uva.nl/praat/

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985.

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

Bradlow, A. R., Pisoni, D., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English/r/and IV: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2223.

Brouwer, S., Mitterer, H., & Huettig, F. (2010). Shadowing reduced speech and alignment. *Journal of the Acoustical Society of America,* 1–14. https://doi.org/10.1121/1.3448022

Brown, H. D. (2015). *Teaching by principles: An interactive approach to language pedagogy* (Vol. 4). Englewood Cliffs, NJ: Prentice Hall Regents.

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.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, *54*(C), 35–50. https://doi.org/10.1016/j.wocn.2015.09.003

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144. https://doi.org/10.1207/s15326969eco0102_2

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dromey, C., & Benson, A. (2003). Effects of concurrent motor, linguistic, or cognitive tasks on speech motor performance. *Journal of Speech, Language, and Hearing Research, 46*(5), 1234–1246.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 33–68). Oxford, UK: Blackwell.

Ellis, N. C. (2009). Optimizing the input: Frequency and in usage-based and form-focused learning. In M. H. Long & C. Doughty (Eds.), *Handbook of language teaching* (pp. 139–158). Oxford, UK: Blackwell.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research.* New York, NY: Routledge.

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187–1199.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York Press

Flege, J. E. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3–28.

Francis, A. L., MacPherson, M. K., Chandrasekaran, B., & Alvar, A. M. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology*, *7*, 263.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldinger, S., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*, 716–722.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds. *Neuropsychologia*, *9*(3), 317–323.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*(4), 611.

Hao, Y. C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168.

Hattori, K. (2010). *Perception and production of English/r/-/l/by adult Japanese speakers* (Doctoral thesis, University College London, UK). Retrieved from http://discovery.ucl.ac.uk/19204/

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America, 119*(5), 3059–3071.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmondsworth, UK: Penguin Books.

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America*, *126*(2), 866. https://doi.org/10.1121/1.3148196

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. *Journal of the Acoustical Society of America*, *118*, 3267.

Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, *46*, 295–349.

Kingston J., & Diehl, R. L. (1994), Phonetic knowledge. *Language, 70,* 419–454.

Kingston, J., & Diehl, R. L. (1995). Intermediate properties in the perception of distinctive feature values. In B. Connell & A. Arvaniti (Eds.), *Phonology and phonetics: Papers in laboratory phonology IV* (pp. 7–27). Cambridge, UK: Cambridge University Press.

Krashen, S. (1981). Bilingual education and second language acquisition theory. *Schooling and language minority students: A theoretical framework* (pp. 51–79). Sacramento, CA: California State Department of Education.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. New York, NY: Addison-Wesley.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.

Kuhl, P. K., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, *55*(4), 306–353. https://doi.org/10.1016/j.cogpsych.2007.01.001

Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. *New Sounds*, *90*, 72–95.

Leussen, V., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology, 6*, 1000.

Libermann, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

Libermann, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, *65*(4), 497–516.

Libermann, A. M., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Libermann, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Libermann, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, *243*(4890), 489–494.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality contrasts: The role of perceptual contrast. *Language, 48*(4), 839–862.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*(7), 1390–1405.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

MacKain, K., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied PsychoLinguistics*, *2*, 369–390.

Maye, J., & Gerken, L. A. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

McClaskey, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of training to a new linguistic contrast in voicing. *Perception & Psychophysics*, *34*(4), 323–330. Retrieved from http://www.springerlink.com/index/Q27V2K3616620482.pdf

McDonough, J., Shaw, C., & Masuhara, H. (2013). *Materials and methods in ELT: A teacher's guide* (3rd ed.). New York, NY: Wiley-Blackwell.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173. https://doi.org/10.1016/j.cognition.2008.08.002

Morrill, T., Baese-Berk, M. M., & Bradlow, A. R. (2016). Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of English. *Proceedings of the International Conference on Speech Prosody, 2016,* 1119–1123.

Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. Language Learning, 68(1), 234–270.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 1–11. https://doi.org/10.1016/j.wocn.2010.12.007

Nunan, D. (2002). Listening in language learning. In J. C. Richards, & W. A. Renandya (Eds.), *Methodology in language teaching* (pp. 238–241). Cambridge, UK: Cambridge University Press.

Nye, P., & Fowler, C. A. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*(1), 63–79.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Pederson, E., & Guion-Anderson, S. (2010). Orienting attention during phonetic training facilitates learning. *The Journal of the Acoustical Society of America*, *127*(2), EL54–EL59.

Pegg, J., Werker, J. F., Ferguson, L., Menn, C A, & Stoel-Gammon, C. (1992). Infant speech perception and phonological acquisition. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 285–311). Timonium, MD: York Press.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461–472.

Polio, C. (2007). A history of input enhancement: Defining an evolving concept. In C. Gascoigne (Ed.). *Assessing the impact of input enhancement in second language education.* Stillwater, OK: New Forums Press.

Prather, J., Okanoya, K., & Bolhuis, J. J. (2017). Brains for birds and babies: Neural parallels between birdsong and speech acquisition. *Neuroscience & Biobehavioral Reviews, 81*(Pt. B), 225–237. https://doi.org/10.1016/j.neubiorev.2016.12.035

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*(6), 608–635.

Samuel, A. G. (2011). The lexicon and phonetic categories: Change is bad, change is necessary. In M. G. Gaskell & P. Zwisterlood (Eds.), *Lexical representation: A multidisciplinary approach*. Berlin, Germany: Mouton de Gruyter.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied PsychoLinguistics*, *3*(03), 243–261. https://doi.org/10.1017/S0142716400001417

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

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.

Sommers, M., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied PsychoLinguistics, 28, 2,* 231–249

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*, 131–145.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition, 119*(1), 131–136.

Tateishi, M. (2013, September 25). *Effects of the use of ultrasound in production training on the perception of English /r/ and /l/ by Native Japanese speakers (*Master's thesis, University of Calgary, Alberta, Canada). Retrieved from https://prism.ucalgary.ca/bitstream/handle/11023/1097/ucalgary_2013_tateishi_miwako.pdf?sequence=2&isAllowed=y

Thorin, J., Sadakata, M., Desain, P., & McQueen, J. M. (2018). Perception and production in interaction during non-native speech category learning. *The Journal of the Acoustical Society of America*, *144*(1), 92–103.

Tremblay, K., Kraus, N., & McGee, T. (1998). The time course of auditory perceptual learning: Neurophysiological changes during speech-sound training. *NeuroReport*, *9*(16), 3556–3560.

Vallabha, G., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *Journal of the Acoustical Society of America*, *116*, 1184.

Vaughn, C. R., Baese-Berk, M. M., & Idemaru, K. (2018). Re-examining phonetic variability in native and non-native speech. *Phonetica*. Advance online publication. https://doi.org/10.1159/00048726

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2/3), 122–144. https://doi.org/10.1159/000107913

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113*(2), 1033–1043.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones: Transfer to production. *Journal of the Acoustical Society of America*, *106*(6), 3649–3658.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.