White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/183598/

Version: Accepted Version

Article:
Lindsay, Shane, Clayards, Meghan, Gennari, Silvia et al. (2022) Plasticity of categories in speech perception and production. Language, Cognition and Neuroscience. ISSN: 2327-3801

https://doi.org/10.1080/23273798.2021.2018471

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Plasticity of categories in speech perception and production

Shane Lindsay[a*], Meghan Clayards[b], Silvia Gennari[c] and M. Gareth Gaskell[c]

[a] Department of Psychology, University of Hull

[b] Department of Linguistics & School of Communication Sciences and Disorders, McGill

[c] Department of Psychology, University of York

Shane Lindsay, Department of Psychology
University of Hull, Hull, HU6 7RX
United Kingdom
Office: (0044) 01482 466539
shane.lindsay@gmail.com

Abstract

While perceptual categories exhibit plasticity following recently heard speech, evidence of effects on production have been mixed. We tested influences of perceptual plasticity on production with an implicit distributional learning paradigm. In Experiment 1 we exposed participants to an unlabeled bimodal distribution of voice onset time (VOT) using bilabial stop consonants, with a longer category boundary than is typical. Participants' perceptual category boundaries shifted towards longer VOT, with a congruent increase in production VOT. Experiment 2 found evidence of perceptual transfer of these shifts to a different speaker and different syllables, and different words in production. Experiment 3 showed no shifts following exposure to a VOT boundary shorter than typical. We conclude that when listeners adjust their perceptual category boundaries, these changes may affect production categories, consistent with models where speech perception and production categories are linked, but with category boundaries influencing the link between perception and production.

Keywords: speech perception; speech production; phonetics; psycholinguistics; statistical learning

Plasticity of categories in speech perception and production

When two people engage in a conversation a peculiar phenomenon may occur where they can end up sounding like each other. This spontaneous tendency to imitate speech has been called phonetic convergence, or accommodation (Babel, 2012; Pardo, 2012). Physiological, dialectal, situational, and social factors can all influence this tendency, though they can sometimes lead to the opposite effect of divergence (Pardo, 2012). Even outside of conversational settings, a tendency to spontaneously imitate heard speech can still be found, such as after hearing ambient natural speech (played over a background loudspeaker) of a different regiolect (Delvaux & Soquet, 2007). It also occurs in shadowing tasks which involve repeating words heard in isolation, leading to productions that tend to resemble the heard speech (Goldinger, 1996; Fowler, Brown, Sabadini, & Weihing, 2003).

One important aspect of this phenomenon is that the influence of a speech percept on production suggests a close correspondence between the systems involved in speech perception and in speech production. In the current work we investigate phonetic convergence in order to further explore the relationship between speech perception and production. It has become increasingly recognized that perceptual categories are inherently plastic, and capable of rapidly altering to facilitate perception of an unfamiliar speaker's accent (Clarke & Garrett, 2004), dialect (Kraljic, Brennan, and Samuel, 2008), or idiosyncratic speech patterns (Norris, McQueen, & Cutler, 2003). In line with theoretical accounts that suggest a close relationship between perception and production categories (e.g., Pickering & Garrod, 2004), it is conceivable that changes to perceptual categories may at least partly underlie the changes that are observed in production. We consider this possibility in the experiments reported in this paper by looking at the short term effects on production following experimentally induced changes in perceptual categories.

A strong link between perception and production is argued for in a number of speech processing models which suggest that production categories rely on goals rooted in perceptual categories, which then act as targets in guiding the articulators towards those auditory goals (Guenther, 1995; Perkell, 2012). The nature of speech categories is considerably debated, but one prominent class of models are distributional theories (Shi, Griffiths, Feldman, & Sanborn, 2010; Smits, Sereno, & Jongman, 2006)[1], where listeners continuously track the frequencies of speech cues and use this distributional information to build models of speech categories. Distributions can be modelled directly by tracking the frequencies of speech cues, potentially using Bayesian inference (e.g., Feldman, Griffiths, & Morgan, 2009; Clayards, 2008; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Norris & McQueen, 2008), or more indirectly in exemplar theories, through episodic storage of memory-rich representations of previous (possibly all) encounters with a word (e.g., Goldinger, 1996; Pierrehumbert, 2001). Distributional theories of speech production often assume a relatively direct link between perception and production, with articulations based on targets derived by sampling from parameterized phonetic category distributions (Kirby, 2010), or by targets sampled from stored exemplars (Goldinger, 2000; Pierrehumbert, 2001, 2002). Given this assumption of how perception and production are linked, distributional models provide a natural account of spontaneous imitation in speech. As perceptual categories are inherently plastic and reflect the phonetic properties of recently experienced speech, with productions that are based on those categories, then productions may converge to properties of the speech of others. Phonetic convergence may be further biased by a weighting towards production targets based on recent experience (Pierrehumbert, 2001).

An alternative perspective to the idea that productions are based on perceptual categories is that the relationship is the other way round, with perceptual categories instead rooted in the production system. This strong link between perception and production is argued for in a variety of models of speech processing, such as the motor theory of speech perception (Liberman & Mattingly, 1985), the direct realist view (Fowler et al., 2003), and articulatory phonology (Browman & Goldstein, 1990). These accounts are consistent with neuroimaging evidence showing that perceptual areas are activated in production (Hickok & Poeppel, 2000), production areas are activated in perception (Wilson, Saygin, Sereno & Iacoboni, 2004), transcranial magnetic stimulation targeting speech production areas can affect phoneme categorization (Meister, Wilson, Deblieck, Wu & Iacoboni, 2007), and perturbation induced motor plasticity in the speech system can affect speech perception (Nasir & Ostry, 2009).

A problem for accounts that claim a strong link between perception and production is that evidence for a close coupling is mixed. A classic example of divergence is Labov's analysis of near-mergers (1994), which documented New Yorkers who could not perceive the difference between *source* and *sauce*, yet produced them differently. Several studies have failed to find any correlation between individuals' perceptual categorization of a voice onset time (VOT) continuum and production of voiced and unvoiced stops (Bailey & Haggard, 1973, 1980; Samuel, 1979)[2]. On the other hand, correlations have been found for the bilabial voicing contrast (Newman, 2003), and there is evidence that correlations between perception and production of voiceless tokens develops over time (Zlatin & Koenigsknecht, 1976). Other evidence for a close relationship comes from work which shows speakers' pronunciation precision is correlated with their ability to discriminate contrasts (Perkell et al., 2004), with speakers who discriminated a contrast more acutely also producing that contrast more distinctly, though this relationship has not always been found (Paliwell, Lindsay, & Ainsworth, 1983). Evidence against motor theories in particular has come from many fronts (for a critical perspective, see Lotto, Hickok, & Holt, 2009), such as the case of anterior aphasics (with lesions to Broca's and related areas) who retained the ability to perceive a VOT contrast normally despite having major problems in producing that contrast, making phonemic and phonetic substitutions (Blumstein, Cooper, Zurif, & Caramazza, 1977).

While there is positive evidence in favor of a link, the many dissociations found between perception and production have led some to argue that the perception-production link is limited, and exists only at an abstract level (e.g., Mitterer & Ernestus, 2008). An additional complication is that some studies have shown that a relationship may exist but it is antagonistic. Baese-Berk (2019) did not find evidence for a clear correlation between perception and production, and found that production of speech during training disrupted perceptual learning in an implicit distributional learning paradigm, and Baese-Berk and Samuel (2016) have also found that producing tokens of novel L2 contrasts led to disruption of perceptual learning. Furthermore, this antagonistic relationship may shift across development; Zamuner, Morin-Lessard, Strahm, and Page (2016) found that production during novel word learning aided perception of novel words in adults, but production led to impairment of perception for 4-6 year olds (Zamuner, Strahm, Morin-Lessard, & Page, 2018).

One avenue for examining the link between perception and production is to see how production is influenced over time by perceptual training, which has been investigated in studies of second-language learning of a novel phonological contrast, but here evidence for a strong link is also mixed. Perceptual training on a novel contrast has been shown to lead to improved

perception and production of that contrast at the group level in the absence of training in production (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). However, Bradlow et al. (1997) failed to find any individual correlations between improvements in perception and production for Japanese speakers' /r/-/l/ contrast. Furthermore, production improvements following perceptual training on an L2 contrast do not always occur (Brosseau-Lapré, Rvachew, Clayards, & Dickson, 2013).

In contrast to studies that attempt to link shifts in perception and production over long time scales such as in second language learning, where underlying mechanisms are hard to pin down, laboratory-based studies allow more immediate assessment of the link between perceptual and production plasticity. One of the few studies that has directly tested whether experimentally induced short-term perceptual plasticity in existing categories is followed by equivalent shifts in production was done by Kraljic et al. (2008). They used a lexical retuning paradigm, where listeners are exposed to ambiguous phonetic tokens that are disambiguated by a lexically biasing context (Norris, McQueen, & Cutler, 2003). Kraljic et al. found that lexical retuning of the ambiguous /s/-/ʃ/ boundary to make /s/ more /ʃ/-like led to perceptual shifts but no change in production of /s/, despite being able to imitate that ambiguous sound after being explicitly asked to. Their results led them to argue that the representations used in perception are separate from those in production, but such a claim is inconclusive on the basis of a single null result with one phonological contrast. This is especially true since positive evidence for a link appears to be dependent on the contrast involved (Newman, 2003). One such case is Gordon and Meyer (1984) who found a link between perception of CV syllables and production of CV syllables when they shared a voicing feature but not when they shared the same place of articulation.

The negative claim of Kraljic et al. also potentially conflicts with other laboratory studies that have focused on plasticity in the production of VOT. Shadowing studies, where participants repeated artificially manipulated hyper-aspirated word-initial voiceless stops (with much longer than typical VOT), found convergence towards longer VOTs in productions (Fowler et al., 2003; Shockley, Sabadini, & Fowler, 2004). These results show the influence of recently heard speech on production targets occurs extremely rapidly (potentially in the order of 300 ms; Tilsen, 2009), which has been taken as offering strong support for episodic exemplar-based accounts that link perception with production. However, one problem for the interpretation of shadowing tasks is they involve deliberate repetition, which may weaken a claim that convergence or imitation is spontaneous and may reflect only an immediate and direct mapping between perception and production without any longer-lasting effects.

A study by Nielsen (2011) that helps to address this issue used an alternative method to the immediate imitation required in shadowing tasks. Participants' voiceless VOT was longer when reading aloud words after hearing those words with longer VOT earlier (word-initial /p/ tokens with +40 ms of natural VOT). One limitation of this work for the present investigation is that as Nielsen and the shadowing studies of VOT discussed above only tested production, it is unknown how any phonetic imitation was linked to learning in the perceptual system. Other work has shown that perceptual learning can be triggered by short term changes in the distributions of speech cues (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Idemaru & Holt, 2011; Maye, Werker, & Gerken, 2002), allowing the modification of existing categories in adults (Clayards et al., 2008). Therefore, recalibration of the voiceless category may have occurred since these studies all involved exposure to many voiceless tokens with hyper-aspirated VOT (mean VOT > 100 ms). Distributional theories of speech perception would predict that exposure

to lengthened VOT could lead participants to update the properties of voiceless categories to match the longer VOT they were hearing, and these altered perceptual categories drove the tendency towards longer mean VOT in production.

In order to better understand the perception-production link and discriminate between different theoretical accounts of the relationship, it is important to address not only whether perceptual plasticity can influence production, but also how this transfer might occur and what information is used to modulate the link. In the current work we sought to create a strong test of a link between perception and production by investigating how experimentally induced perceptual shifts would influence production. We manipulated VOT based on the evidence it is susceptible to rapid perceptual (e.g., Clarke & Luce, 2005; Sumner 2011) and production shifts (e.g., Nielsen, 2011). Our method involved measuring baselines for both perception and production, and then exposing participants to a novel VOT continuum which they also categorized. This allowed the testing of shifts in perception both during (allowing assessment of the time course of perceptual shifts) and after exposure, and whether any perceptual shifts occurred with production shifts. An important consideration was to use a more implicit methodology than many other studies. Like Nielsen (2011), our procedure avoided the direct imitation used in shadowing experiments, and was reinforced by using a novel method to induce shifts in perceptual categories which had a high degree of implicitness. Whereas the lexical retuning paradigm used by Kraljic et al. (2008) and others used a top-down signal of the lexical category label to modify existing perceptual categories, we relied on statistical learning in the absence of disambiguating information to operate as a form of unsupervised learning.

Motivated by the evidence that the speech system is sensitive to the probability distributions of acoustic cues (e.g., Clayards et al., 2008; Kleinschmidt, Raizada, & Jaeger, 2015), we exposed participants to an artificially constructed bimodal distribution of a word-initial bilabial VOT continuum (/b/-/p/) with the aim of shifting the boundary between categories. Implicit learning of a manipulated distribution of VOT has now been shown to lead to shifts in categorization boundaries in a number of studies investigating plasticity in speech categorization (Kleinschmidt, Raizada, & Jaeger, 2015; Munson, 2011) as well as other continua (e.g., Colby, Clayards, & Baum, 2018; Llompart & Reinisch, 2018). Unlike those experiments on VOT production where only the voiceless category was manipulated and tested (cf. Fowler et al., 2003; Nielsen, 2011; Shockley et al., 2003), using a full continuum of VOT allows manipulation of both elements of the contrast. A further advantage was our manipulation of VOT was more subtle than those experiments investigating learning effects on production that relied on large and uniform shifts in VOT using hyper-aspiration (i.e., > 100 ms VOT; Fowler et al., 2003; Nielsen, 2011; Shockley et al., 2004).

To illustrate our manipulation, a natural occurring VOT distribution and the artificial exposure continuum we designed for Experiment 1 is illustrated in Figure 1. Panel A shows how the boundary between categories for our artificial distribution (grey bars) would be longer than the naturally occurring boundary between the voiced /b/ category (on the left) and the voiceless /p/ category (on the right). Participants heard tokens in minimal pairs (e.g., beach/peach) with no feedback. Due to the absence of a top-down learning signal to determine category membership for ambiguous tokens, any perceptual changes must rely on learning the statistical properties of the artificial continuum. Assuming a speech category system that is plastic and sensitive to the statistical features of recent speech experience, we predicted that participants would modify their categorization behavior such that category boundaries would more closely match the boundary of

the artificial distribution. Measuring perception and production allowed testing whether any perceptual changes (indicated by category boundary shifts) would influence production.

Based on accounts which argue for a separation between perception and production (e.g., Kraljic et al., 2008), one possible outcome of the manipulation was that it could affect perception but not production, which would be seen through phonetic convergence in perception but the absence of any evidence of changes in production categories. Alternatively, A further possibility was that if categories in perception and production are strongly linked (e.g., Pickering & Garrod, 2004), or even shared to some extent (e.g., Liberman & Mattingly, 1985), shifts in perceptual categories should be linked with immediate changes in production. If production changes occurred along with perceptual changes, a following question was how information in the probability distribution of acoustic cues influenced perceptual category modification, and subsequent effects on production. We addressed this question by manipulating the relationship between our artificial exposure continuum and the natural category structure. An important property of the artificial distribution used in Experiment 1 (shown in Panel B of Figure 1) was that it involved a mean-shift manipulation as well as a boundary-shift: the mean of the artificial voiced equivalent category was longer than typical of English (indicated by the rightwards arrow in Panel B), whereas the mean of the voiceless equivalent distribution was shorter than typical of English (indicated by the leftwards arrow). If productions were directed towards the most likely or most representative categories of heard speech (i.e., the means of the distributions) then post-exposure productions should be longer for voiced tokens, but shorter for voiceless tokens.

An alternative possibility was that rather than relying on distribution means, information signaling the category boundary would be particularly important in determining category structure and modulating the perception-production link, as would be predicted by perceptual categorization decision-bound models (e.g. Ashby & Perrin, 1993) which emphasize the role of the boundary in speech categorization (Liberman, Harris, Hoffman, & Griffith, 1957; Maddox & Chandrasekaran, 2014; Smits, Sereno & Jongman, 2006).

Through exposure to the lengthened boundary, and a subsequent shift in their categorization boundary towards longer VOT, participants may have shifted their productions towards longer VOT. In this case a voiceless category shift towards longer VOT would be in the opposite direction to the mean shift, and what would be predicted by distributional models.

A further question we wanted to address in the current work was the abstractness of the representations involved. Exemplar models in particular have been argued to operate at the lexical level, with learning effects more tied to specific lexical items (Goldinger, 2000). In Experiment 1, we initially addressed this question by using syllables (/ba/-/pa/) in our pre- and post-exposure perceptual tasks, and using words with different syllables in our exposure phase (e.g., beach/peach). If learning was tied to particular syllables or words we would not expect to have found that perceptual learning to transfer to different syllables. However, as lexical retuning paradigms have shown perceptual learning can transfer to new words containing a recalibrated phoneme (McQueen, Norris, and Cutler, 2006), if bottom-up statistical learning operates in a similar fashion we would also expect transfer to new syllables, indicating learning at a sub-syllabic level (cf. Munson, 2011).

Experiment 1

Method

Participants

Twenty-three participants from the University of York were tested. All were native British English speakers without visual or auditory impairments. Participants received course credit or were paid £6 for participation. Sample size was based on Clayards et al. (2008) which had an n = 24 using an unsupervised learning paradigm similar to used here and found evidence for perceptual categorization shifts.

Design

The experiment comprised three phases, which are illustrated in Figure 2. In the pre-exposure phase participants first took part in the baseline production task, followed by the baseline syllable categorization task. In the exposure phase participants heard our artificial VOT distribution three times over three blocks and were asked to categorize each token without any feedback. In the post-exposure phase, participants repeated the production and syllable categorization tasks. In our analysis, we looked for evidence of adaptation to the exposure distribution shown by a change in the categorization function across the three blocks, measured as the proportion of /p/ responses. In the production task we looked to see if VOT of productions (with /b/ and /p/ initial words) changed in comparing pre vs. post-exposure. The syllable categorization task looked at whether there was a shift in the categorization functions comparing pre vs. post exposure responses.

Materials and Stimulus Construction

Stimuli in the exposure phase were made up from three minimal pairs (beach/peach, beak/peak, & beat/peat). Recordings of the voiced words heard in the exposure phase (/b/ words) and the syllables /ba/ and /pa/ were taken from a female native British English speaker. A total of 27 examples of /ba/, 30 examples of /pa/, and 7 examples of each /b/ word (beach, beak, beat) were recorded. A clear token of each was selected. We constructed the word and syllable voicing continua by splicing the release and aspiration from a single recording of /pa/ onto the vowel and, in the case of the words, the vowel and final consonant of a recording with an initial /b/. Thirteen splice points were selected for each stimulus at rising zero crossings, with the first splice point beginning at 2 ms after sound onset. Subsequent splice points were selected at approximately 5 ms intervals (see McMurray & Aslin, 2005, for a similar procedure). Due to the need to splice at zero crossings to avoid auditory artifacts our VOT values reported below are approximate, but were always within 2 ms of the target. For example, when we report a VOT of 20 ms, the true VOT would lie between 18 ms and 22 ms. The /p/ stimuli were then spliced onto the corresponding /b/ stimuli at the splice points, for each of the three minimal pairs and the syllable. This created continua with gradually increasing VOT and decreasing duration of the subsequent vowel, ranging from 10 to 70 ms of VOT, with the total length remaining constant (with slight variation due to zero crossing constraint). Note that natural voicing continua do involve multidimensional cues. While VOT is by the far the most important, other cues such as F1 are also used by listeners to discriminate contrasts, and are weighted particularly highly in ambiguous regions of the VOT continua where VOT is a less informative cue. While it could be argued that isolating a single cue from a multi-dimensional speech signal would reduce the similarity of our stimuli to natural speech, it ensures listeners' sensitivity to the distributional properties of this cue to this a cue as driver for category change (e.g., Samuel, 1979, 1982).

The bimodal distribution in the exposure phase we designed for Experiment 1 is illustrated in Figure 1. The VOT mean of the voiced VOT category (/b/) was 25 ms and 55 ms for the voiceless /p/ equivalent (*SD* = 8.1 ms in both cases). For comparison, in Figure 1 a VOT distribution taken from our pre-exposure production data has been overlaid and scaled to match the artificial continuum. It is known that population VOT distributions are slightly positively skewed (Allen & Miller, 2001) but for the present purposes a Gaussian distribution is a reasonable approximation. Our intended exposure distribution, shown in Figure 1 by the grey bars, was made up of 168 tokens. This consisted of 56 tokens for each of the three minimal pairs, with 24 tokens at the two peaks of the distribution (made of eight tokens from each minimal pair), and just three tokens (one for each minimal pair) at the far ends of the distribution shown in Figure 1 (10 ms and 70 ms). The entire distribution was heard by participants three times, once in each of the three blocks. We designed our artificial distribution to deviate from a normal distribution slightly in order to have the frequency at each VOT step divisible by three, to allow equal numbers of our three minimal pairs to be presented in each of the three blocks of the exposure task. Compared with a natural VOT continuum, our distribution was more symmetrical and the /b/-/p/ boundary was shifted towards longer VOT. The size of the shift was intended to make any perceptual or production effect as clear as possible, while also allowing the two peaks in the distribution to be close enough to natural /b/ and /p/ peaks to ensure that they were accepted as belonging to those categories. The vertical lines on Panel A of Figure 1 shows the position of the boundaries. For comparison with the new artificial boundary (which maximally separates the peaks of our artificial distribution), we show the category boundary derived from the mean VOT of maximal ambiguity in participants' categorization from the first block of the exposure phase (i.e., where categorization was 50% /b/ and 50% /p/). The size of the shift was around 10 ms longer compared with this approximation of the existing boundary. Panel B illustrates that the left-hand peak of our artificial VOT distribution was approximately 12 ms longer than a natural peak of /b/ initial productions (in the same direction as the boundary shift). The right-hand peak of our artificial VOT distribution was in the opposite direction, 8 ms shorter than the peak of a typical /p/ distribution, and the distribution itself was narrower with less kurtosis.

Due to a script coding error, it transpired there was a deviation from the intended frequencies which only affected the first two steps of the VOT continuum in Experiment 1. This deviation is shown in the frequencies of Table 1 alongside the intended distribution. The beat-peat items had the intended distribution, but beach/peach and beak/peak items had different exposures at 10 ms and 15 ms than intended, with a differing pattern for 8 participants and another one for 15 participants. While not exactly what we had originally planned the heard distribution still had the properties we intended, which was a bimodal distribution with minimal exposures in the tails, two peaks of high frequency density, and a sparse region in between the peaks. The deviations led to the same means of the left side of the distribution (steps 1 to 6) and the same standard deviation for the 8 participant group and a very small change in standard deviation for the 15 participants (20.7 instead of the original 21.3). Therefore we do not believe these slight deviations impacts any conclusions we would want to make from the exposure used.

Procedure

The experiment lasted approximately an hour, with the distribution categorization task taking approximately 45 minutes. Participants were permitted to take short rests in between

tasks. Instructions for each task were presented on screen. They were given an overview of the procedure and were told that it was an experiment on perception and production. Participants were instructed to respond as quickly and accurately as possible. Once the experiment had started, each of the three phases (pre-exposure, exposure, post-exposure) followed sequentially on from each other without any experimenter intervention, which meant participants did not hear any linguistic inputs other than experimental stimuli. Stimuli were heard at a comfortable listening level over high quality headphones, and productions were recorded using Beyerdynamic DT 234 Pro headsets. The experiment was run using DMDX (K. I. Forster & J. C. Forster, 2003) on individual participants in a quiet room, and responses for the categorization tasks were recorded using a USB joypad.

Production (pre- and post-exposure).

Words were presented one at a time in the middle of the screen in lower case font, and participants had 1200 ms to respond, after which the next word appeared. Participants were instructed to speak the visually presented word into the headset microphone naturally and normally. Each of the six words used in the exposure phase (beach/peach, beak/peak, & beat/peat) was presented three times over three blocks, in an order randomized per participant. The items also included four filler words, presented three times in each phase, which differed pre-exposure (beer/peer & bees/peas) and post-exposure (beep/peep & beast/pieced). Four practice trials were provided in both phases with words not seen or heard in any other part of the experiment.

Syllable Categorization (pre- and post-exposure).

Participants heard 78 stimuli taken from a /ba/-/pa/ continuum (6 tokens for each of 13 steps). On each trial participants saw a fixation cross on screen, and then 800 ms later the letters “ba” or “pa” appeared on the left or right side of a centrally located fixation point, followed by the spoken stimuli 900 ms later. Trials were terminated by a response, or timed out after 4000 ms. Participants had to indicate which target syllable they felt the speaker was saying by pressing the corresponding button on the joypad (left button for the syllable on the left of the fixation point, right button for the syllable on the right). The position on the screen was counterbalanced, such that “ba” appeared on either side of the fixation point for an equal amount of trials as its corresponding “pa”. This was done to prevent participants associating either buttons on the joypad systematically with a particular syllable. Order of token presentation was randomized per participant. They were instructed to respond as quickly and accurately as possible and always make a response. There were four practice trials.

Exposure.

Participants saw a centrally located fixation cross on screen, and then 800 ms later both items from one of the three minimal pairs appeared in lower case font on the left or right side of the fixation point (e.g., “beach” on the left, “peach” on the right), followed by the spoken word 900 ms later. The side of presentation was counterbalanced in a similar way to the syllable categorization task. Trials were terminated by a response, or timed out after 4000 ms from the onset of the carrier phrase. Within each of the three blocks participants had to classify each stimulus as one of the two words using the joypad, with the left or right button corresponding with the position of the word on the screen. There was no feedback. Participants categorized 504

tokens in total (168 tokens in each block). The order of each presentation was randomized for each participant for each block. Participants were given an opportunity to take a short break between blocks.

Results

Categorization in the Exposure Phase

Scripts and data for our results can be found at https://osf.io/9a4gm/.

Trials which timed out and no response recorded represented 0.1% of the data in Experiment 1, 1.1% in Experiment 2 and 1.4% in Experiment 3.

Figure 3 shows the mean categorization function for all participants during the three blocks of the exposure phase. In assessing learning, we compared categorization across the three blocks, expecting a shift in participants' categorization function across blocks, with later blocks showing more /b/ responses, consistent with a rightwards shift in the boundary between unvoiced and voiced. We used mixed-effects modeling with the statistical package R, using the lme4 package (Bates, 2007; R Development Core Team, 2008). For our categorization tasks our response was binary (/b/ or /p/) therefore we used a logistic link functions (Jaeger, 2008), with /b/ coded as 0 and /p/ coded as 1. Starting from a simple model with VOT step (13 levels) as a fixed effect and random intercepts for participant we conducted model selection using likelihood ratio tests, keeping only factors that significantly affected the fit of the model and allowed the model to converge successfully. Model selection led to the inclusion of block (3 blocks), and item (the 3 minimal pairs) as fixed effects, and the slopes of VOT step were allowed to vary with the random participant intercepts. This model assumes that participants vary in their categorization functions across the range of VOT. Item contrasts were sum coded (deviation coding, using base r), which allows interpretation of the intercept of the model as the mean response across items, and VOT was included as a continuous variable centered so that the intercept was the mean response at our mid-point in the continuum (step 7; ~40 ms, coded as 0), allowing for easier interpretation of the coefficients. We used categorization responses during the second block as the reference level (0) with dummy coding to allow orthogonal contrasts with block 1 and block 3. The final model used was glmer(formula = outcome ~ 1 + VOT + block + item + (VOT|subid)).

This model showed a significant learning effect, indicated by a rightwards shift in the categorization function in Figure 3 across blocks, with participants producing significantly fewer /p/ and more /b/ responses in the second block than the first block (b = 0.423, SE = 0.09, z = 4.88, $p < .001$), but with no significant change between the second and third blocks (b = -0.046, SE = .088, z = -0.53, p = .6). The increase in /b/ responses indicates a boundary shift towards longer VOT. In other words, participants changed their criterion for assigning /b/ or /p/ responses such that in order to categorize the same token as /p/ in the second block they needed that token to have a longer VOT compared with the first block. This shift appeared to occur relatively early on during exposure. Across our models with categorization as the dependent variable we found highly significant effects of VOT step, as would be expected given the shape of the categorisation functions with /p/ responses increasing as the VOT cue increased. Full mixed model outputs for our results are included in the OSF repository.

Syllable Categorization

We examined whether exposure affected categorization of the syllable continuum in the post-exposure phase, by comparing it with responses in the pre-exposure phase.

RT's Responses greater than 4000 ms and shorter than 400 ms in the categorization tasks in Experiment 1 were removed from the data, representing 0.1% of the data in Experiment 1, 1.0% in Experiment 2 and 1.4% in Experiment 3. Due to software malfunction, one participant's data was lost from the post-exposure phase, but they were kept in the mixed-effects model.

Categorization functions for each phase are illustrated in Figure 33. The model selection procedure followed the categorization task. Starting from a model that included VOT (centered) as a fixed effect and random intercepts for participants led to additional inclusion of fixed effects of phase (pre or post-exposure; contrast coded with pre as the reference level) and random slopes for VOT by participants intercepts. The final model used was glmer(formula = outcome ~ 1 + VOT + phase + (VOT|subid)). This model showed a nonsignificant effect of phase (b = - 0.225, SE = 0.13, z = -1.74, p = .083), with a greater numerical tendency towards /b/ categorization post-exposure compared with pre-exposure.

Production

Coding of VOT was completed by research assistants with phonetic training but naïve to the purposes of the experiment, and was done by careful listening to the recordings combined with visual inspection of spectrograms and waveforms under high magnification with the Praat software package (Boersma, 2001). Trials in which the participant made an error, did not respond or the VOT was not discernable made up 3.3% of the data.

We used two separate mixed-effect models to analyze the exposure effect for word initial /b/ and /p/ VOT for the three item-pairs heard in the distribution categorization phase, with model implementation similar to above but using a Gaussian link function instead of the logistic link function as our dependent variable here was continuous. We report p-values generated from Markov chain Monte Carlo sampling (Baayen, Davidson & Bates, 2008). Model selection starting from a model with participants as random intercepts led to the inclusion of word (sum coded) and phase (dummy coded, with pre-exposure as the reference level) as fixed effects. The final model used was lmer(formula = VOT ~ 1 + item + phase + (1|subid)). Following exposure, VOT production of /b/ initial words were significantly longer (b = 2.63, SE = .5, z = 5.36, p < .001), as were /p/ initial words (b = 4.67, SE = 1.35, z = 3.45, p < .001). Means can be found in Table 2 and pre- and post-exposure distributions frequency polygons with accompanying empirical cumulative distribution plots are presented in Figure 4.

Discussion

Experiment 1 demonstrated that distributional properties of a phonetic continuum can quickly induce perceptual shifts even in the absence of lexical feedback, with categorization boundaries shifted towards longer VOT over the course of exposure. In an additional perceptual categorization task, we found a nonsignificant effect on syllable categorization post-exposure. Most noteworthy in our results was that despite a nonsignificant effect in the syllable categorization task, plasticity induced by exposure to the distribution transferred to production, with longer productions of both /b/ and /p/ initial words.

Even though our participants heard VOTs that were on average shorter than usual in English for the voiceless category, participants produced VOTs that were longer in duration for

these items following exposure to our artificial distribution. This finding is opposite to that predicted by models of phonetic convergence that use knowledge about exemplars or peaks in distributions to influence or generate targets for the production system. One interpretation of the longer voiceless VOT in productions is that the location of the implicit category boundary between voiced and voiceless items in exposure (i.e., the relative scarcity of syllables with VOT of 30 to 40 ms) was the crucial modulator of productions. This shifted boundary (in comparison with the typical boundary for this contrast) may have had the effect of pushing VOT for the /p/ category on the right hand tail further away from the boundary. That the boundary was particularly important is highlighted in decision-bound models of categorization (e.g., Liberman, Harris, Hoffman, & Griffith, 1957; Smits et al., 2006).

Following the perceptual learning shown in the exposure phase, the syllable categorization task did not show reliable evidence for learning effects transferring to different syllables in a separate task. Clear evidence for a null result would be informative about the level of abstraction necessary to explain these effects, indicating that learning was specific to those items, but the results of Experiment 1 do not clearly discriminate between a null effect or a real effect.

One aspect of the exposure phase used in Experiment 1 (and used in Experiment 2) is that it has some similarity to a selective adaptation paradigm (Eimas & Corbitt, 1973). Selective adaptation occurs when rapid repeated exposure to tokens taken from one end of a perceptual continuum leads to reduction in categorization of that end of the continuum (i.e., hearing lots of clear /b/ tokens leads to a reduced likelihood to categorize an ambiguous token as a /b/). Most selective adaptation paradigms involve presentation of identical clear tokens from just one end of a continuum. In comparison with a natural continuum (see Figure 1), our artificial /p/ tokens for Experiments 1 and 2 overlap considerably with a natural /p/ distribution. However, the peak of the artificial /b/ distribution falls close to the natural boundary between /b/ and /p/, which is the region of maximal ambiguity. This could result in listeners perceiving many clear /p/ tokens, comparatively fewer clear /b/s, and many ambiguous tokens. This might have resulted in selective adaptation to /p/, leading to the observed effect of more /b/ responses overall.

At first glance, our design might appear quite different from selective adaptation paradigms, which often involve several iterations of passive listening to a large number of identical clear tokens presented in very rapid succession (e.g., 100 times within a minute, Sawusch & Pisoni, 1976), immediately followed by a categorization task. However, selective adaptation has been found in a variety of experimental paradigms, such as following presentation of tokens in connected speech (Rudnicky & Cole, 1977). While adaptation appears to weaken over time when tested over different time intervals (Sharf & Ohde, 1981), it still can be present in some participants up to 28 minutes after exposure (Sharf & Ohde, 1981), and recent unpublished evidence indicates adaptation effects were largely intact after 25 minutes and diminished but still significant after 6 hours (Samuel & Dumay, 2021). We therefore cannot rule out the possibility that selective adaptation had an influence on our results in perception, though we highlight the work of Kleinschmidt and Jaeger (2016), who argue that selective adaptation is best seen as a form of distributional learning.

Tests of selective adaptation transfers to production (“perceptuomotor adaptation”) using syllable-initial voiceless stops have found shortened VOT in productions of those stops (Cooper, 1974; Cooper & Lauritsen, 1974; Cooper & Nager, 1975; Jamieson & Cheesman, 1987; though

see Summerfield, Bailey, & Erickson, 1980). Such an effect makes sense in the context of selective adaptation, where adaptation to a voiceless token is thought to lead to an increase in voiced categorization due to VOT values being perceived as being lower than they actually are, and hence production VOT should also be lowered (Cooper, 1979, p. 120). However, the findings witnessed here were in the opposite direction to previous findings of perceptuomotor adaptation, with lengthened VOT rather than shortened VOT. This provides a reason to doubt that selective adaptation was the driver for production changes, though selective adaptation remains viable as an explanation for our perception changes.

Experiment 2

In Experiment 2, we had several goals. We wanted to replicate the findings main findings of Experiment 1 of significant evidence for perceptual learning in exposure and transfer to production, and to help clarify effects of generalization in the syllable categorization task, where the results were inconclusive. Figure 4

In extending the findings, we aimed to further explore the question of generalization. In Experiment 1, exposure words were heard in isolation and spoken by the same female talker heard in the syllable categorization task, with the burst and aspiration portions cross-spliced so that they were identical for both words and syllables. In Experiment 2, we tested whether any perceptual shifts transfer to the perception of a new speaker. If so, it would be indicative that the representations involved are not tied to a specific speaker, and hence are somewhat abstract. In order to test the level of specificity, the stimuli from Experiment 2 included the same burst and aspiration portions from Experiment 1 in the exposure phase to ensure that the key phonetic cues were identical across experiments and between phases. However, in Experiment 2 the bursts and aspirations in the exposure phase were spliced onto words spoken by a different female talker to the one heard in the syllable categorization phase (cf. Eisner & McQueen, 2005; Kraljic & Samuel, 2006), and were embedded in semantically neutral sentences (e.g., “now choose the one that is beach/peach”) spoken by the same speaker as the exposure words. As an additional benefit of this design change, the use of sentences also more closely approximates natural listening situations, allowing us to observe if plasticity still occurs when participants are exposed to connected speech.

A further question regarding generalization was whether our exposure-induced changes in production operated at a lexical (e.g., linked to a lexical item such as “beach”) or sub-lexical level (e.g., linked to a syllable /bi/, a phoneme /b/ a feature [±voice], or a context-dependent allophone). Whilst there are several studies that suggest perceptual learning operates at a sub-lexical level (e.g., McQueen, Norris & Cutler, 2006; Mitterer, Reinisch, & McQueen, 2018), less is known about the level of representations involved in production shifts. To test this aspect of generalization we looked at whether three minimal pairs not heard in the exposure phase would differ in their production VOT pre- and post-exposure, compared with the three minimal pairs used in the exposure phase. Given some evidence for learning at a sub-lexical level in the syllable categorization task (with transfer to different syllables), if perception and production are linked we should expect production shifts to also occur with different words. This would indicate a sub-lexical locus for such shifts, as suggested by the findings of Nielsen (2011), who found transfer (sub-phonemic) in production to /k/ after hearing longer /p/’s.

Experiment 2 also allowed us to rule out the possibility that our shift in production was due to a slower speaking rate in the post-test by examining word length for the participants’

productions. It is known that VOT is influenced by speaking rate, with longer VOT for slower speaking rates (i.e. longer utterances, Allen & Miller, 1999; Miller, Green, & Reeves, 1986). Due to a technical problem in Experiment 1 the end parts of some words were not recorded, therefore we ran this analysis in Experiment 2 only.

Method

Participants

Twenty-three participants from the University of York were tested. All were native English speakers without visual or auditory impairments, who had not taken part in Experiment 1. Participants received course credit or were paid £6 for participation.

Design and Procedure

The experiment was very similar to Experiment 1, but with some key differences. The syllables for the /b/-/p/ continuum (pre- and post-exposure) were identical to those of Experiment 1. For the exposure task we again constructed words by splicing the release and aspiration from a word recorded with an initial /p/ onto the vowel and final consonant of a word recorded with an initial /b/. This time, however, the /p/ initial words were taken from the recordings of the speaker in Experiment 1 while the /b/ initial words were taken from a second female native British English speaker. In this way, we kept the acoustics of the burst and release identical to Experiment 1 and across the pre-exposure, exposure and post-exposure phases, while the indexical information in the subsequent vowel, indicating talker identity, was new. Recordings were made by the new speaker reading from a list of three sentences two times, which contained a neutral (semantically un-biasing) carrier phrase without word-initial stop consonants (“this one is”, “now choose the one that is”, “the answer she chose was”) and a /b/ target word 500 ms after the final word of the carrier phrase (beach, beak, beat; 9 combinations in total). The counts for the exposure distribution are shown in Table 1.

All recordings were then filtered to remove energy below 75 Hz, which reduces noise that can occur due to AC current and makes identification of zero crossings more difficult. The pitches of /b/ words were manipulated using the PSOLA (Pitch Synchronous Overlap and Add) algorithm in Praat (Boersma, 2001) to match the pitch of the /p/ word of the minimal pair (e.g., pitch of “beach” matched to “peach”) to reduce pitch differences between the voices of the speakers, and stimuli were normalized to match loudness. During the experiment, one version of each of the three sentences was heard, followed by words from our VOT continuum. As a result of pilot testing our sentence stimuli, we found our categorization functions were slightly more biased towards /p/ responses compared with Experiment 1. In order to counteract this we shifted the distribution by one step towards shorter VOT, which meant our distribution started at 5 ms and peaked for the /b/ at 20 ms (cf. 10 ms start and 25 ms peak in Experiment 1).

To help familiarize participants with the voice used in the exposure phase, and therefore maximize their awareness that it was a different voice from the pre- and post-exposure syllable categorization tasks, we included a combination of on-screen instructions and spoken instructions over the headphones. We excluded word-initial stop consonants in our spoken instructions, and minimized the use of stop consonants in other word positions. These instructions were almost identical to Experiment 1, with mostly the same text used in Experiment 1 in written format with some information transferred to a spoken format, with a few word

changes to avoid stop consonants. At the beginning of the task participants were shown an overview of the experiment on screen, informing them that the speaker of the instructions was the same native English speaker of the stimuli in the exposure phase, and that this was a different native English speaker from the stimuli in the pre- and post-exposure syllable categorization tasks.

The syllable categorization task was the same as Experiment 1, using exactly the same stimuli, but with an additional reminder that the speaker was native English and was a different speaker to the one who spoke the instructions. Participants were reminded in the post-exposure phase that the speaker of the syllables was different to the speaker of the stimuli in the exposure phase.

Like Experiment 1, in the production task we presented the same minimal pairs heard in the exposure phase (beach/peach, beak/peak, & beat/peat). In order to test generalization across productions in Experiment 2, we used an additional six items in both pre- and post-exposure (bees/peas, beer/peer, beep/peep) that differed from those items heard during exposure.

Results

Categorization in the Exposure Phase

Figure 5 shows the average categorization responses across the three blocks of exposure in Experiment 2. We used the same model specification as Experiment 1 but with the design of this study including carrier sentences, model selection led to the inclusion of sentence (3 levels) as an additional sum coded fixed effect in the model along with block and item as in Experiment 1. Like Experiment 1, we found a significant shift early on in the exposure task, with a significant increase in /b/ responses from the first to the second block ($b = 0.606$, $SE = 0.094$, $z = 6.22$, $p < .001$). In Experiment 1 there was no significant further increase between the second and third blocks, but in Experiment 2 this later categorization shift was also statistically reliable ($b = -0.242$, $SE = 0.01$, $z = -2.42$, $p = .015$).

Syllable Categorization

Figure 5 shows the average categorization responses for the pre- and post-exposure syllable categorization task in Experiment 2. The same structured mixed-effects model as Experiment 1 revealed a significant shift in participants' categorization functions, with significantly more /b/ responses following exposure, ($b = -0.471$, $SE = .146$, $z = -3.22$, $p < .01$), in comparison to only a marginal effect in the same direction in Experiment 1. This confirms that the perceptual shifts seen here were not word, syllable or speaker specific.

Production

Figure 6 shows the production distributions and cumulative distributions of VOTs pre- and post-exposure for Experiment 2. Missing data constituted 1.6% of the total available in Experiment 2. Using the same model structure as in Experiment 1, for the three item-pairs heard in the exposure phase we found that VOT of /b/ productions was slightly longer following exposure but unlike Experiment 1 this was not statistically significant ($b = .83$, $SE = .465$, $t = 1.8$, $p = .072$). However, as found in Experiment 1, there was a significant shift in /p/ initial productions towards longer VOT ($b = 4.94$, $SE = 1.62$, $t = 3.06$, $p < .01$). In order to see if effects transferred to other words, we used a separate model for the three item pairs that were not heard

in the exposure phase (bees/peas, beer/peer, beep/peep). The results mirrored the above findings, with no significant VOT change for /b/ initial words, (t =.53), but significantly longer VOT for /p/ initial words (b = 6.486, SE = 1.727, z = 3.76, p < .001) [3].

Since longer productions (i.e. slower speaking rate) are correlated with longer VOT, particularly for /p/ productions, we needed to ensure that the longer VOT found in post-exposure were not simply due to longer productions post-exposure compared with pre-exposure. The lengths of productions were measured (from the onset of the burst to the end of the final consonant) by research assistants naive to the experimental manipulations, and by a different research assistant to the coder of the VOT, with 1.8% of the data not included due to unclear recordings. Pre-exposure productions actually were slightly longer (/b/ M = 417 ms, SD = 46 ms; /p/ M = 456 ms; SD = 48 ms) rather than shorter compared with post-exposure productions (/b/ M = 413 ms, SD = 43 ms; /p/ M = 453 ms; SD = 51 ms). A paired t-test did not show a significant difference between pre- and post-exposure lengths (p > .39), nor did we find any significant interactions between phase and length when we added this to mixed-effects models for /b/ and /p/ productions (p > .44) . This indicates that our results of longer production VOT post-exposure was not due to longer productions overall.

Discussion

Like Experiment 1, we found large categorization shifts in the exposure phase. This perceptual learning effect with words now showed reliable evidence of generalization to shifts in categorization of different initial syllables post-exposure. While we must be cautious in the absence of a significant effect in the perceptual categorization task in Experiment 1, learning effects with different syllables suggest that the statistical learning observed is neither specific to lexical or syllabic-level representations (since the vowels differed between test syllables and exposure words). We demonstrated that effects in perception transferred to a different speaker with a different syllable to that used in exposure, indicating that plasticity in phonetic categories was neither word, syllable nor speaker specific. Note, however, that we did not test whether it was specific to the acoustics of the release and aspiration used in the construction of all stimuli. The effects of exposure on perception also influenced productions of VOT. Effects appeared weaker than in Experiment 1, with the shift statistically significant only for the /p/ category. A further finding of Experiment 2 was that the production effect was also observed for words not heard in the exposure phase, suggesting a sub-lexical locus of the effect in production.

In Experiment 2 we used connected speech in exposure, compared with single words in Experiment 1, but in both cases our production task used single words. This meant that the exposure matched the production type in Experiment 1 but not in Experiment 2. As conversational speech is typically faster (and hence has shorter VOTs), the expected distributional statistics of VOT in normal everyday speech should be shifted towards shorter VOT compared with expected distributions of words heard in isolation (at least for voiceless tokens, as unvoiced stops tend to be less affected by speaking rate; Allen & Miller, 1999; Miller, et al., 1986). Despite this difference in speaking styles, and a distribution of VOTs shorter than what might be expected for that speaking style, we still found indications of lengthened productions in Experiment 2. This provides further evidence that what is learned in one speech style transfers to another speech style. It should be noted, however, that this discrepancy between experiments is somewhat mitigated by our use of a relatively large gap between the last word of the carrier sentence and the target word, in addition to the fact that it was the final word in the

sentence, which contained new and unpredictable information, and therefore would be expected to be spoken more slowly and clearly than had it occurred earlier in the sentence.

One issue for the interpretation of both the previous experiments is that it is conceivable that shifts were due to some confounding variable or variables, which led to a tendency towards more /b/ responses in categorization during exposure and/or longer VOT in production, producing a bias unrelated to our particular exposure distribution. For example, one possibility suggested by the work of Baese-Berk and Goldrick (2009), is that presentation of minimal pairs together encouraged hyperarticulation of the voiceless member (i.e., lengthened VOT of /p/). Potentially this effect (a tendency to speak more clearly) then increased across the experiment with sustained exposure to minimal pairs.

In Experiment 3, we aimed to rule out such the possibility of lengthened VOTs due to reasons other than exposure to the altered distribution.

Experiment 3

Experiment 3 was designed to test whether the post-exposure effects found in Experiments 1 and 2 were a direct result of the properties of the inputs received in the exposure phase. As well as serving as a control experiment, Experiment 3 allowed the opportunity to explore how other manipulations of the exposure distribution could influence perception and production. In Experiment 3 we used the same shaped bimodal distribution as Experiment 2 but shifted it towards shorter VOT, which is illustrated in Figure 7. This led to pre-voicing (negative VOT, which is produced with glottal pulsing before the aspiration) for the left-hand tail of the distribution, and a shift in the category boundary towards the voiced end of the VOT continuum. While voiced stops in English typically have positive VOT, some pre-lag voicing does occur (Lisker & Abramson, 1964), and even relatively large amounts of pre-voicing does not normally sound abnormal. A further advantage of this shorter VOT shift is that it arguably provides a better match to what might be expected with exposure to running speech (a distribution with shorter VOTs compared with isolated speech), particularly for the voiceless category which typically has a shorter VOT with conversational speech than in clear isolated speech (Allen & Miller, 1999).

One possible consequence of our manipulation in Experiment 3 was that as a result of distributional learning, participants would shift their perceptual boundary towards shorter VOT, and this could in turn lead to shorter VOT in production. However, existing results have shown difficultly in producing perceptual (VanDam, 2007) and production shifts (Nielsen, 2011) towards shorter VOT in comparison to shifts towards longer VOT. Either way, for the purposes of a control experiment, a reversal of the shift found in the previous studies or a null result would indicate that it was the particular nature of our exposure distribution which caused perceptual and production shifts in the previous experiments.

Method

Participants

Twenty-six participants from the University of York were tested. All were native English speakers without visual or auditory impairments, who had not taken part in Experiment 1. Participants received course credit or were paid £6 for participation.

Design and Stimulus Construction

The stimuli and procedure were identical to that of Experiment 2 apart from two differences. In Experiments 1 and 2, the boundary of our exposure distribution was shifted to have a longer VOT than a typical English bilabial voicing continuum (Lisker & Abramson, 1964). In Experiment 3 we used an exposure distribution (see Figure 7) where the peaks of our artificial /b/ and /p/ categories, and the boundary point between them, was shifted to have shorter VOT (approximately 10 ms) than what is typical in English, including some tokens with pre-voicing. In order to construct stimuli with pre-voicing, a 65 ms sample of pre-voicing was taken from a naturally pre-voiced stop produced by the speaker from Experiment 1. This was shortened to create each of the pre-voiced steps using the PSOLA algorithm in Praat (Boersma, 2001). The intensity was scaled to have a peak of 73 dB. Peak intensity of the following vowel was 80 dB on average. Each of these pre-voiced steps was then pasted before the 0 ms token from the VOT continuum.

A second difference from the previous experiments was in the syllable categorization task (which as before was indicated to be a different speaker). Given that we were presenting shorter VOT during exposure in Experiment 3, and that the boundary of the syllable categorization task in Experiment 2 was already in the lower-VOT end of our tested range (about 5 steps into our 13 step continuum; see Figure 7) there was a possibility of the boundary being pushed too close to the low-VOT end of the categorization continuum in Experiment 3. Consequently, we altered the VOT continuum for the pre- and post-exposure tests to spread from 0 ms to 60 ms (compared with 10 ms to 70 ms in Experiments 1 and 2), based on the assumption that the boundary point would be roughly similar (around 30-35ms) and that point would fall mid-way between the range of VOT values used.

Results

Categorization in the Exposure Phase

Figure 8 shows the categorization functions for the three blocks of the exposure phase. The mixed-effects model (with same structure as Experiment 2) revealed no significant difference in categorization between the first or second block, (z = .59) nor between the second and third block (z = *-.04*). To provide the strongest test of learning, we also used a contrast between the first (as reference level) and third block, which was also not significant (z = -0.63).

Syllable Categorization

Due to an intermittent software error in recording, 11 participants had a random trial missing on the pre-exposure task or the post-exposure task, or both (4 participants), leading to a loss of data of 0.47% overall. Missing trials made up an additional 0.5% of the data removed.

Figure 8 shows the categorization functions for the pre- and post-exposure syllables. The mixed-effects model (with the same structure as Experiment 1 and 2) showed no significant difference in syllable categorization comparing pre- and post-exposure (b = *-0.211*, *SE* = .119, z = -1.78, p = .08).

Production

There was 2.8% of the data missing in Experiment 3. For /b/ and /p/ productions, for both sets of word pairs (heard or not heard in exposure), the effect of exposure was not significant (all $t < 1.33$Figure 9, along with the means inTable 4).

Discussion

In Experiment 3, where the artificial exposure distribution was shifted towards shorter VOT, we found no change in phonetic categorization in the exposure phase. While caution is always needed in the face of null findings, these null results provide some help in interpreting the results of the previous experiments. In particular, it offers support for the conclusion that the results of the two previous experiments arose due to the particular properties of their exposure distribution rather than some other factor. Compared with Experiment 2 in particular, the otherwise identical manipulation of Experiment 3 did not bring about the same effects, though we should note a qualitatively similar pattern in the syllable categorization task.

As we found shifts in perception and production in the earlier experiments, it is perhaps unsurprising that we did not find significant evidence for shifts in the syllable categorization and production tasks post-exposure, given no clear effect of perceptual learning effects during exposure. Consistent with our lack of evidence for a perceptual shift in Experiment 3, VanDam (2007) compared perceptual categorization of word initial voiceless stops after 5 days of exposure to either shortened (80% of typical /p/) or lengthened (180%) VOT for /p/ initial words. He found a significant perceptual categorization shift in the category boundary following exposure to lengthened VOT, but no boundary shift following exposure to shortened VOT. In this case a possible factor was the magnitude of the shift of shorter VOT was smaller than the VOT lengthening manipulation, due to a need to preserve natural sounding speech (low VOT for /p/ initial words would make them sound like /b/'s).

However, in the perceptual domain the lack of perceptual categorization shifts with shortened VOT is not predicted by distributional models in which perceptual categorization would be influenced by recently heard speech. It also contrasts with other findings with manipulations of lower VOT (Clarke & Luce, 2005; Kleinschmidt & Jaeger, 2016; Sumner, 2011), where shifts in categorization have occurred when the voiced category was altered to pre-lag voicing, and voiceless tokens were moved into the voiced range. In fact, there are a number of differences across studies that could explain the discrepancy in findings. Perhaps the most noteworthy is the size of their manipulations. Clarke and Luce used /d/ tokens having a mean of -44 ms (range -18 to -68 ms) and /t/ tokens with a mean of 30 ms (range 22 to 35 ms). Sumner used /b/ initial tokens with pre-voicing as great at -70 ms and /p/ initial tokens as low as 0 ms. They used a number of different exposure conditions (our exposure was most similar to their random condition), but across conditions the average for /b/ was -35 ms and for /p/ was +35 ms. Kleinschmidt and Jaeger (2016) used mean shifts of approximately -10 ms for /b/ and +30 ms for /p/. These are more extreme divergences of perceptual experience compared with typically encountered speech than the shift used in Experiment 3, especially when compared with our voiced category (/b/ M = 5 ms, range = -10 to 20 ms; /p/ M = 35 ms, range = 20 to 50 ms), and with a larger separation between the voiced and voiceless means. A further complication is that it is not clear how sensitive English speakers are to pre-voicing, and whether it is perceived categorically. It remains possible that had we exposed participants to a more radically shifted distribution for both voiced and voiceless categories it would have led to evidence for perceptual shifts, and consequently, evidence for effects on production.

There was another factor that differed between our study compared with Sumner and Clarke and Luce, which was their use of lexical feedback with non-minimal pairs or biasing sentences. This may have enhanced the possibility of learning compared with the unsupervised learning used here (Goudbeek, Swingley, & Smits, 2009; though see Kleinschmidt et al., 2016), and allows more radical distributional manipulations (where substantial lowering VOT of a voiceless category member brings it into the voiced range) than possible with a statistical learning paradigm without feedback.

Our lack of significant evidence for a production shift is in accord with Nielsen's (2011) finding that exposure to /p/ initial words with much lengthened VOT led to longer VOT in production, but exposure to shortened VOT (/p/s with -40 ms of typical VOT, M = 31 ms) did not lead to any significant production changes. She argued that the lack of a production shift was due to a need for contrast preservation, as shortening VOT of the voiceless category without a change in the voiced category would reduce the phonemic contrast between categories. This might be avoided by the production system if one of the goals in production is to maximize contrast for listener benefit (Flemming, 2004; Lindblom, 1990). However, as perceptual shifts were not tested, the alternative possibility exists that if production shifts occur in the presence of perceptual shifts (as we found in Experiment 1 and 2), Nielsen's manipulation of shorter VOT may not have been sufficient to induce a perceptual shift.

General Discussion

Our results indicate that perceptual plasticity in category boundaries as a result of listening to spoken words in speech can generalize to perception of different syllables from different talkers (Experiment 2), and to production (Experiment 1 and 2), even when cues to category boundaries are only signaled implicitly by distributional information. In Experiment 1 and 2, over the course of 504 unlabeled stimulus presentations in the exposure phase, participants showed a shift towards longer VOT in their perceptual category boundaries congruent with the location of a dip in the bimodal exposure distribution, as a consequence of exposure to single words (Experiment 1) or connected speech (Experiment 2). Thus, participants were able to shift their perceptual boundaries on the basis of distributional properties of fine-grained phonetic detail (cf. Munson, 2011). This categorization shift with the training materials also influenced more general perception of VOT. Categorization generalized to a /ba/-/pa/ continuum, showing a congruent boundary shift after exposure in Experiment 2. Following these perceptual category shifts, we found generalization to production of the key consonants, with longer VOT in the post-test than the pre-test. In Experiment 1 this effect was found for both /b/ and /p/'s, but only significantly for /p/'s in Experiment 2. In contrast, following exposure to shortened VOT in Experiment 3 there was lack of evidence perceptual learning or shifts in production.

The results of Experiment 1 and 2 show perceptual shifts co-occurring with production shifts. This is in contrast to claims that plasticity in perception do not impact production (Kraljic et al., 2008), and gives suggestive evidence for a link between perception and production. However, this link would appear somewhat different from what is described in models which assume that productions reflect exemplars or the central tendencies of recently heard speech, with the production system relying on targets from the perceptual system (e.g., Guenther, 1995; Kirby, 2010; Perkell, 2012; Pierrehumbert, 2001). These models do not easily explain two key aspects of our data. First, in Experiment 3 (in which all exposure VOTs were shortened) we found no learning effects. The second, and perhaps noteworthy aspect of the production data,

was that the direction of the production shift for the voiceless /p/ category. The direction of a predicted shift influenced by the most commonly encountered speech is shown in the right panel B in Figure 1, which compares a natural /p/ distribution with our artificial distribution. Instead, the results of Experiments 1 and 2 (shown in Figure 4 and Figure 8 respectively) show a production shift in the opposite direction. This pattern of results is clearly counter to a distributional model. Motor theories, which assume that the targets of perception are articulatory gestures (Liberman & Mattingly, 1985) rather than auditory properties, could also be interpreted to make similar predictions. If production categories were activated and consequently altered during the perceptual exposure phase, they should tend to shift towards toward the most frequently occurring speech.

Instead of mirroring of the exposure data, one possibility is that the low frequency of exemplars in the boundary region between the peaks may have been influential in inducing changes between pre- and post-exposure VOT in perception and production. For the /p/ consonants after exposure, participants tended to favor longer VOTs that were relatively far from the exposure boundary. Boundary or decision-bound models of speech perception have a long history (Liberman, Harris, Hoffman, & Griffith, 1957), and more recent boundary models of categorization can account well for phonetic categorization data compared with other theories (Smits et al., 2006), and have been applied to speech category learning (Maddox & Chandrasekaran, 2014). Our results are consistent with a model in which there is plasticity in the boundaries between perceptual categories, and that listeners could use information about the parts of the continuum that are sparsely populated as evidence for a boundary, which may drive production changes to maintain contrast.

This follows Nielsen (2011) who argued that speakers' mirroring of speech is constrained by the avoidance of productions that reduce contrast (i.e., speech near a boundary), to help explain why she found no change in VOT in production of voiceless plosives after hearing voiceless tokens with shortened VOT, a similar result to our Experiment 3. On the bilabial voicing continuum, this contrast would suggest avoidance of long /b/'s and/or short /p/'s. In our case, while the tendency we found towards longer /b/ VOT would be in violation of contrast preservation mechanisms on its own, this shift occurred in combination with a shift towards longer /p/'s, which means the relative distance between categories would be preserved and contrast maintained. That speakers are consistent in maintaining distance between categories was found by Samuel (1979), who showed a reliable correlation between speakers' productions of voiced and voiceless stops; examination of our data found this same pattern[4]. While our results are consistent with the idea that contrast preservation is a factor that influences how perceptual shifts are expressed in productions, the lack of perceptual shifts in Experiment 3 and the lack of measuring such shifts in Nielsen (2011) leaves the importance of its role inconclusive.

As well as contrast preservation, other articulatory factors could modulate the link between perception and production. While the perceptual system can adapt categories in response to a speaker's contrasts which depart from typically encountered English speech, such as pre-lag voicing (Clarke & Luce, 2005; Sumner, 2011), lack of articulatory practice within that range of VOT values may limit how those perceptual changes influence productions. Even in the case of typical English VOT, there are articulatory factors which limit the upper range of voiced productions (Kessinger & Blumstein, 1997). These articulatory factors have been used to explain why cross linguistically the short lag VOT category (the voiced category in English) is less variable than the long lag VOT category (the voiceless category in English), and why in slower

(Miller et al., 1986) and clearer speech (Smiljanić & Bradlow, 2005) the English voiceless category tends to change more (towards longer VOT and more variability with slower speech), with comparatively little change in the English voiced category. If speakers already are used to producing a wide range of VOT values for the English voiceless category, this pre-existing flexibility may make this category more labile and responsive to fluctuations in perception compared to the more fixed voiced category. In our case, we did find evidence for a shift for the less variable voiced category, but this was only reliable in Experiment 1 and was smaller than the shift in the voiceless category. This asymmetry has also been found in perceptuomotor adaptation studies with voicing contrasts, where selective adaptation in perception transfers to production have only been found for the voiceless category (Cooper, 1974; Cooper & Lauritsen, 1974; Cooper & Nager, 1975; Jamieson & Cheesman, 1987).

While our results support a link between plasticity in perception and production, there has been debate as to how strong this link is due to the body of evidence that speech perception and production are not directly linked (e.g., Bailey & Haggard, 1973, 1980; Blumstein et al., 1977; Bradlow et al., 1997; Kraljic et al., 2008). Some of this negative evidence can be explained as arising from methodological factors (e.g., Newman, 2003), but in regards to the general question of how strong the link is, frameworks in which the production system relies on auditory-temporal goals give a possible account of the link. These frameworks emphasize the use of complimentary representations in the production system (somatosensory goals and articulatory representations; Guenther, 1995; Perkell, 2012). While auditory-temporal goals appear necessary for the development of the production system in infancy, in later life the production system does not have to rely on the auditory-temporal goals (and the reliance on goals can differ depending on the contrast; Perkell, 2012). This means, for example, that learning could occur in the production system without it necessarily being linked to perceptual representations (as found in Baese-Berk, 2019). Whether or not the use of auditory-temporal goals in production (such as using perceptual boundaries to bias productions) is best characterized as a strong or relatively weak link, this theory provides an explanation for the positive evidence for a relationship as well as providing an account of how they can be dissociated.

In comparing our work to the negative results for a link between perception and production, one unresolved question is why Kraljic et al. (2008) found a null result for effects on production following plasticity induced by lexical retuning of existing speech categories. There are a number of differences between studies which may help explain their null result and our more positive findings, which include the type of learning and the phonological contrast involved. While we do not see any compelling reasons why top-down lexically determined learning versus more bottom-up statistical learning should lead to differential effects for transfer to production, different parameters of exposure may have had an impact. Kraljic et al. only attempted to find a production shift for one phoneme in one direction, with a perceptual shift of /s/ towards the lower spectral frequency /ʃ/ category, which would have reduced the contrast between categories. One possibility is that other manipulations of perceptual learning would have been more likely to lead to production changes, such as a perceptual shift of /s/ away from /ʃ/ towards higher frequencies. This would increase the contrast between the categories, and studies of clearly articulated sibilants have found a tendency to increase the frequency of /s/ productions away from /ʃ/ towards higher frequencies for which there is no upper bound or bordering category (Maniwa, Jongman, & Wade, 2009), with little evidence for changes in /ʃ/. If production changes are constrained by contrast preservation mechanisms and influenced by

articulatory factors which determine category flexibility, this manipulation may promote the likelihood of a production shift.

A further potentially important factor which may influence the link between perception and production and the discrepancies in finding a relationship are the phonemes involved (e.g., Newman, 2003). In comparing stop consonants and sibilants (used by Kraljic et al., 2008), one difference may be the abstractness of the representations. We found evidence for generalization in perception across different syllables and speakers in Experiment 2. Other work on talker specific effects in perceptual plasticity have had varied results; while lexical retuning of sibilants has been shown to be tied more to individual speakers (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), our results are consistent with lexically tuned boundary shifts on a voicing continuum which have not been found to be talker specific (Kraljic & Samuel, 2006; Munson, 2011). One explanation for these differences is that in the case of vowels and fricatives, contrasts are cued spectrally and contain considerable speaker specific information with regard to vocal tract size, leading to vowels contributing more to talker recognition than consonants (Owren & Cardillo, 2006). In contrast, stop consonants may provide limited details about speaker identity, as they contain mostly temporal information with little spectral information (restricted to transient bursts and aspiration). Consequently, one speculative possibility to explaining the discrepancy between our results and Kraljic et al.'s (2008) is that information which is able to generalize in perception (i.e., timing in stop consonants) might be more likely to show automatic generalization to production, and to different place of articulation in production (Nielsen, 2011).

Our results showing that learning effects are not tied to specific lexical items, speakers, or syllables has been taken to indicate pre-lexical abstraction in phonological categories (McQueen et al., 2006; Mitterer, Reinisch, & McQueen, 2018), which has been argued to be better explained in abstractionist models compared to exemplar models (Cutler, Eisner, McQueen, & Norris, 2010). Findings of generalization as well as evidence for episodic effects (e.g., Goldinger, 1996) has led to increased support for hybrid models in which abstract phonological categories and more specific phonetically detailed representations co-exist (e.g., Pierrehumbert, 2002; Pisoni & Levi, 2007). It should be acknowledged that the presence of abstraction is not necessarily an overwhelming argument against distributional models. In parametric theories categories are stored as parametric abstractions from speech experience (such as Gaussian mixtures; Smits et al., 2006). Abstraction also does have a fundamental role in exemplar models (Walsh, Mobius, Wade, & Schutze, 2010), though occurring at a recognition stage (e.g., Goldinger, 1996), as central tendencies emerge when accessing long term memory. However, our evidence for generalization was not robust across experiments, therefore caution should be applied for strong conclusions regarding the level of abstraction mediating effects in this study.

Studies of phonetic convergence show that listeners' tendency to match (or mismatch) that of interlocutors is influenced by a variety of complex linguistic and paralinguistic factors, and can serve a number of purposes (Pardo, 2012). This ability relies on a system that is able to track fine phonetic detail in speech and use such information as targets for production (though see Mitterer & Ernestus, 2008). It has been argued that the ability of the speech perception system to store this detailed statistical information occurs primarily for the purpose of facilitating imitation for social functions (Poeppel, Idsardi, & Van Wassenhove, 2008). However, we suggest the opposite, in that the functional employment of phonetic convergence is made possible through co-opting the mechanisms that are responsible for the formation and modification of speech categories in perception and production.

While we have focused on the use of statistical learning of distributions of phonetic cues (particularly boundary information) as the principle factor in determining phonetic convergence in our experiments, there are a range of other factors which can determine how convergence manifests. For example, the degree to which our participants identified with the speaker may modulate the degree of convergence shown (Babel, 2012). While we have not considered such factors here, we do not suggest they are unimportant. Disentangling these influences, and the relative reliance on phonetically detailed and more abstract representations (Mitterer & Ernestus, 2008) remains an important goal for further research. Additionally, such factors will need to be incorporated in subsequent models (see Johnson, 2006, for one attempt to include social information in an exemplar model). A further outstanding question is that of individual differences in plasticity. While these studies were focused on understanding group level effects, future work showing links between perception and production at the level of individual participants would provide stronger evidence for a causal link between perceptual and production shifts, and some caution should be applied in the absence of such evidence.

Our results demonstrate that statistical learning of shifted category boundaries in the absence of lexically disambiguating information speech categories can perceptual influence speech categories and transfer to production. In future application of this technique, more work is needed to determine to what extent the results found here – including rapid categorization shifts in the absence of lexical disambiguation, generalization to a new talker, and transfers to production - are true of speech categories in general and to what extent these characteristics differ across phonemic contrasts (see Colby et al., 2018, for work in this direction), acoustic cues, both uni and multi-dimensional, and underlie individual level differences. In addition, we should acknowledge that some effects were not large and consistently reliable across experiments, such as in the syllable categorization task and the /b/ production effect, and may not be robust to different analytic choices. As is often the case, it would be worth replicating these effects future work with larger samples. In summary, our results show that perceptual plasticity in category boundaries as a result of listening to words on their own or in continuous speech can generalize to perception of different talkers and to production, even when cues to category boundaries are only signaled implicitly using distributional information. While these results support a link between perception and production categories, our results are not consistent with models in which productions are based on a direct mapping to perceptual categories which are updated continuously by tracking exemplars or means and variances. Instead, it appears that listeners are sensitive to statistical information signaling the boundaries between categories, and this boundary information appears particularly important in the modulation of productions.

Acknowledgements

The research was supported by a UK ESRC research grant RES-063-27-0061 awarded to M.G.G. We gratefully acknowledge the assistance of Sunil Naran, Claire Blackmore, Elaine Tham, Henness Wong and Mila Baer with data collection and coding, and Arty Samuel and Rachel M. Theodore for their comments.

References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*, 2031.

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Attention, Perception, & Psychophysics, 63*(5), 798-810.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review,* 95, 124–150.

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.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*, 177-189.

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, 1-25.

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.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and cognitive processes*, *24*(4), 527-554.

Bailey, P. J., & Haggard, M. P. (1973). Perception and production: Some correlations on voicing of an initial stop. *Language and Speech, 16*, 189–195.

Bailey, P. J., & Haggard, M. P. (1980). Perception-production relations in the voicing contrast for initial stops in 3-year-olds. *Phonetica, 37*, 377–396.

Boersma, P. P. G. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-18.

Blumstein, S., Cooper, W.E., Zurif, E.B., & Carmazza, A. (1977). The perception and production of Voice-Onset Time in aphasia. *Neuropsychologia*, *15*(3), 371-372.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299–2310.

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

Brosseau-Lapré, F., Rvachew, S., Clayards, M., Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics, 34*(3), 419-441.

Browman, C. & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics,* 18, 299-320.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, *54*(2), 421–428.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*, 3647.

Clarke, C. M., & Luce, P. A. (2005). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. In C. T. McLennan, P. A. Luce, G. Mauner & J. Charles-Luce (Eds.), *University at Buffalo Working Papers on Language and Perception, 2* (pp. 362-366).

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Speech perception reflects optimal use of probabilistic speech cues. *Cognition, 108*, 804-809.

Clayards, M. (2008). *The ideal listener: making optimal use of acoustic-phonetic cues for word recognition*. PhD thesis, University of Rochester.

Colby, S., Clayards, M., & Baum, S. (2018). The Role of Lexical Status and Individual Differences for Perceptual Learning in Younger and Older Adults. *Journal of Speech, Language, and Hearing Research*, *61*(8), 1855-1874.

Cooper, W. (1974). Perceptuomotor adaptation to a speech feature. *Attention, Perception, & Psychophysics, 16*(2), 229–234.

Cooper, W.E. (1979). *Speech perception and production: Studies in Selective Adaptation.* Norwood, NJ: Ablex Publishing Co.

Cooper, W. E., Ebert, R. R., & Cole, R. A. (1976). Speech perception and production of the consonant cluster (st*). Journal of Experimental Psychology: Human Perception and Performance, 2*(1), 105-114.

Cooper, W E., & Lauritsen, M. R. (1974). Feature processing in the perception and production of speech. *Nature, 252*(5479), 121–123.

Cooper, W. E., & Nager, R. M. (1975). Perceptuo-motor adaptation to speech: an analysis of bisyllabic utterances and a neural model. *The Journal of the Acoustical Society of America, 58,* 256-266.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.) *Laboratory Phonology* (Vol. 10, pp. 91–111). Berlin, Germany: de Gruyte.

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica, 64*(2-3), 145-173.

Eimas, P.D., & Corbit, J.D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*, 99-109.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics, 67*(2), 224-238.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In Hayes, B., Kirchner, R., & Steriade, D. (Eds.), *The Phonetic Bases of Markedness*. Cambridge University Press: Cambridge.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments and Computers, 35,* 116-124.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396-413.

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

Goldinger, S.D. (2000). The role of perceptual episodes in lexical processing. In A. Cutler, J. M. McQueen, & R. Zondervan (Eds.), *Proceedings of SWAP: Spoken Word Access Processes* (pp. 155–159). Nijmegen: Max-Planck Institute for Psycholinguistics.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1913.

Guenther, F. H., (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102*(3), 594.621.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*, 131–138.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. Journal of Phonetics, 24, 485–499.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, *69*(3), 505-528.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics, 25*(2), 143-168.

Kleinschmidt, D. F., Raizada, R. D., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P.Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1129–1134). Austin, TX: Cognitive Science Society.

Kleinschmidt, D. F., & Jaeger, T. F. (2016a). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review, 23*(3), 678-691.

Kleinschmidt, D. F., & Jaeger, T. F. (2016b). What do you expect from an unfamiliar talker? In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.

Kirby, J. P. (2010). *Cue selection and category restructuring in sound change*. PhD thesis, University of California.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141-178.

Kraljic, T., & Samuel, A. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects and speech processing. *Cognition, 107,* 54-81.

Labov, W. (1994). *Principles of linguistic change: Internal* factors (Vol. 1). Cambridge, MA: Blackwell.

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

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Lindblom, B. (1990). Explaining variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer Academic.

Lisker, L., & Abramson, A. S. (1964). Cross-language study of voicing in initial stops. *Word, 20*, 384-422.

Llompart, M., & Reinisch, E. (2018) Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition, 1-16.* Online First.

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110-114.

Maddox, W. T., & Chandrasekaran, B. (2014). Tests of a dual-systems model of speech category learning. *Bilingualism, 17*(4), 709–728.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America, 125,* 3962-3973.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, *17*(19), 1692-1696.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113-1126.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15-B26.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1-3), 106-115.

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.

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language, 98,* 77-92.

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing.* Doctoral dissertation, University of Iowa (ProQuest No. 3494079).

Nasir, S. M., & Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences*, *106*(48), 20470 -20475.

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *The Journal of the Acoustical Society of America, 113,* 2850-2860.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*(2), 132-142.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357-395.

Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America, 119*, 1727-1739.

Pardo, J. S. (2012). Reflections on Phonetic Convergence: Speech Perception does not Mirror Speech Production. *Language and Linguistics Compass*, *6*(12), 753-767.

Paliwal, K. K., Lindsay, D., & Ainsworth, W. A. (1983). Correlation between production and perception of English vowels. *Journal of Phonetics, 11*(1), 77-83.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics, 25*(5), 382–407.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, *116*(4), 2338-2344.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-190.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological studies in language, 45*, 137-158.

Pierrehumbert, J.B. (2002). Word-specific phonetics, In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101–139). Berlin: Mouton.

Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 3–18). Oxford, UK: Oxford University Press

Poeppel, D., Idsardi, W. J., & Van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1071-1086.

Rudnicky, A. I., & Cole, R. A. (1977). Adaptation produced by connected speech. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 51-61.

Samuel, A.G. (1979). *Speech is specialized, not special.* PhD thesis, University of California, San Diego. Dissertation Abstracts International, 40-08B.

Samuel, A. G. (1982). Phonetic prototypes. *Attention, Perception, & Psychophysics, 31*(4), 307–314.

Samuel, A.G., & Dumay, N. (2021). Auditory selective adaptation moment by moment, at multiple timescales. *Journal of Experimental Psychology: Human Perception and Performance, 47*(4), 596-615.

Sawusch, J. R., & Pisoni, D. B. (1976). Response organization in selective adaptation to speech sounds. *Attention, Perception, & Psychophysics*, *20*(6), 413-418.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

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

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

Summerfield, Q., Bailey, P. J., & Erickson, D. (1980). A note on perceptuo-motor adaptation of speech. *Journal of Phonetics, 8,* 499.

Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 32*(3), 733.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118*(3 Pt 1), 1677–1688.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276–296.

VanDam, M. (2007). *Plasticity of phonological categories*. PhD thesis, Indiana University.

Zamuner, T. S., Strahm, S., Morin-Lessard, E., & Page, M. P. (2018). Reverse production effect: Children recognize novel words better when they are heard rather than produced. *Developmental Science*, *21*(4), e12636.

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. P. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, *89*, 55-67.

Zlatin, M. A., & Koenigsknecht, R. A. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech, Language and Hearing Research, 19*(1), 93.

Walsh, M., Mobius, B., Wade, T., & Schutze, H. (2010). Multi-level exemplar theory. *Cognitive Science*, *34*, 537–582.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*,701–702.

Footnotes

[1] We group exemplar and parametric models by describing them as distributional theories following Smits et al., as the predictions of these models have been difficult to distinguish empirically (Smits et al., 2006), with Smits et al. describing the key difference as whether distributions are modelled non-parametrically (in exemplar models) or parametrically. It has also been suggested that exemplar models may be a way to implement parametric models (Shi, Griffiths, Feldman, & Sanborn, 2010).

[2] VOT is a temporal cue to voicing in word-initial position in English, which we focus on in this paper. It measures the latency between the release burst of stop consonants and the onset of voicing, characterized by glottal pulses that lead to periodic high-amplitude waveforms associated with a subsequent vowel. In English the bilabial VOT contrast consists of two overlapping clusters of short-lag VOT voiced stops (/b/), where burst release and voicing occur closely in time, and long-lag VOT voiceless stops (/p/), where burst release and voicing are more separated in time.

[3] Although we had originally not intended to look at generalization in the design of Experiment 1, at the suggestion of a reviewer who noted due to the similarities between the item (labial stops with similar vowels), we also ran separate models for the four item pairs not heard in the exposure phase in Experiment 1 which differed for the post-exposure task and originally intended as fillers (beer/peer & bees/peas pre-exposure, beep/peep & beast/pieced post exposure). Here we also found that VOT for productions of /b/ words was longer post-exposure (M = 15.1 ms, SD = 6.8 ms) compared with pre-exposure (M = 13.5 ms, SD = 5.8 ms; b = 1.61, SE = 0.80, z = 2.01, p = .046). However, here the significant increase in VOT seen for the repeated items in Experiment 1 was not seen in the VOT for the non-repeated /p/ productions comparing pre (M = 67.5 ms, SD = 14.0 ms) and post-exposure (M = 67.7 ms, SD = 14.8 ms; b = 1.83, SE = 2.52, z = 0.73, p = .47).

[4] Across experiments, we found a significant correlation between voiced and voiceless productions, $r(70) = .34$, $p < .01$.

Table 1

Frequency counts for exposure distributions per participant for each experiment. Exp. 1^{15} refers to 15 participants who received this distribution and Exp 1^{8} to the other 8 participants.

	Exp. 1^{15}	Exp 1^{8}	Exp. Intended	Exp. 2	Exp. 3
-10					9
-5					27
0					54
5				9	72
10	15	27	9	27	54
15	21	9	27	54	27
20	54	54	54	72	18
25	72	72	72	54	27
30	54	54	54	27	54
35	27	27	27	18	72
40	18	18	18	27	54
45	27	27	27	54	27
50	54	54	54	72	9
55	72	72	72	54	
60	54	54	54	27	
65	27	27	27	9	
70	9	9	9		
Total	504	504		504	504

Table 2
Means for productions of VOT (in ms) in Experiment 1. Means and SD generated from the ezStats function from r package ez.

	Mean	Std. Dev.
/b/ initial words		
pre-exposure	13.0	5.5
post-exposure	15.6	6.9
/p/ initial words		
pre-exposure	62.7	14.2
post-exposure	67.6	15.3

Table 3

Means for productions of VOT (in ms) in Experiment 2. Means and SD generated from the ezStats function from r package ez.

		Mean	Std. Dev.
Repeated	/b/		
	pre-exposure	14.4	6.0
	post-exposure	15.2	5.4
	/p/		
	pre-exposure	74.6	19.2
	post-exposure	79.5	21.2
Unrepeated	/b/		
	pre-exposure	15.1	6.5
	post-exposure	15.4	4.8
	/p/		
	pre-exposure	78.2	19.1
	post-exposure	84.8	20.2

Table 4

Means for productions of VOT (in ms) in Experiment 3. Means and SD generated from the ezStats function from r package ez.

		Mean	Std. Dev.
Repeated	/b/		
	pre-exposure	17.5	4.2
	post-exposure	17.5	3.7
	/p/		
	pre-exposure	78.9	14
	post-exposure	78.1	15.3
Unrepeated	/b/		
	pre-exposure	17.7	3.4
	post-exposure	18	3.4
	/p/		
	pre-exposure	80.6	14.5
	post-exposure	82.8	15.4

Figure 1. Histogram showing the artificial VOT distribution used in the exposure phase in Experiment 1. In order to allow comparison with a naturally occurring VOT distribution a frequency polygon is overlaid showing a VOT distribution taken from the pre-exposure productions in Experiment 1. Panel A (top) shows the direction of shift in the boundary towards longer VOT when comparing natural and artificial distributions. Panel B (bottom) shows the comparison of longer VOT for the /b/ category but shorter VOT for the /p/ category. As the natural VOT distribution involved a different than our actual distribution the scale of has been adjusted to facilitate comparison with the artificial distribution for illustrative purposes.

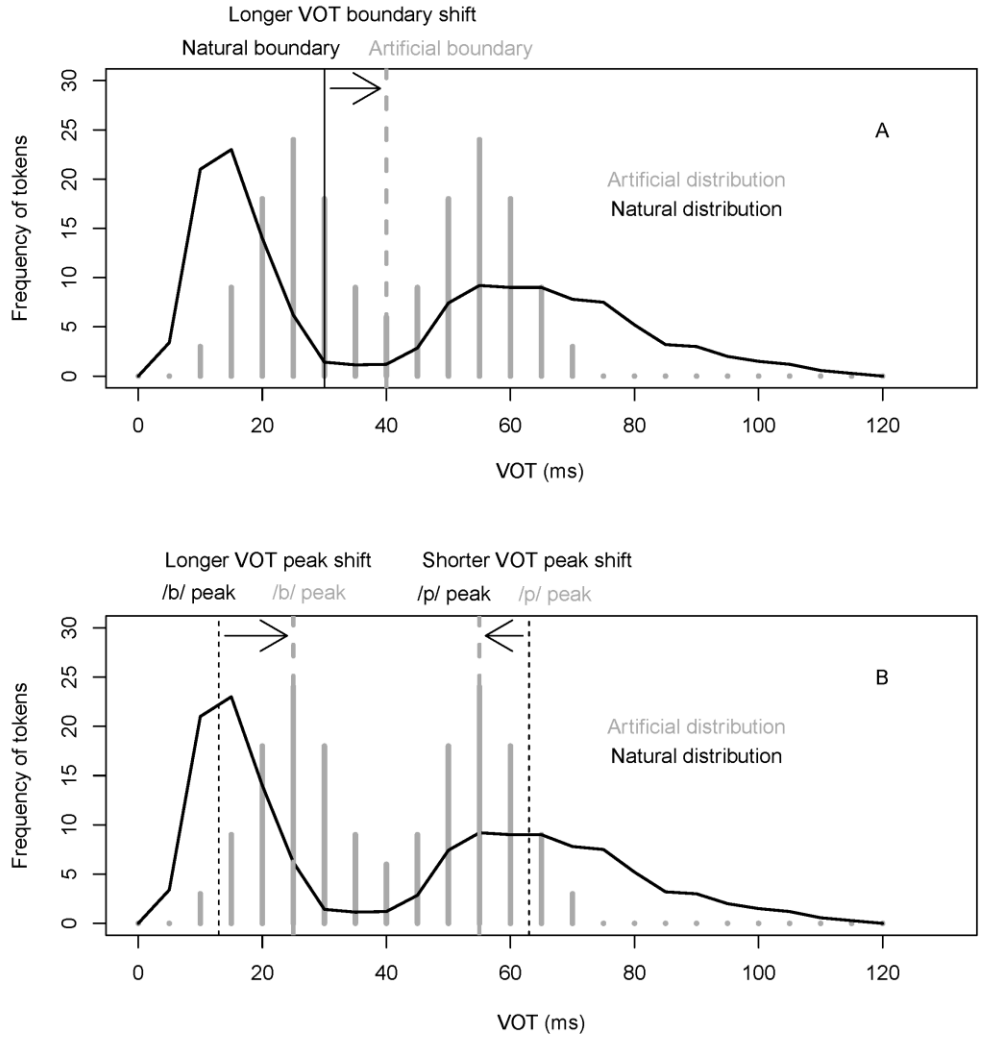


Figure 2. Diagram showing experimental design.

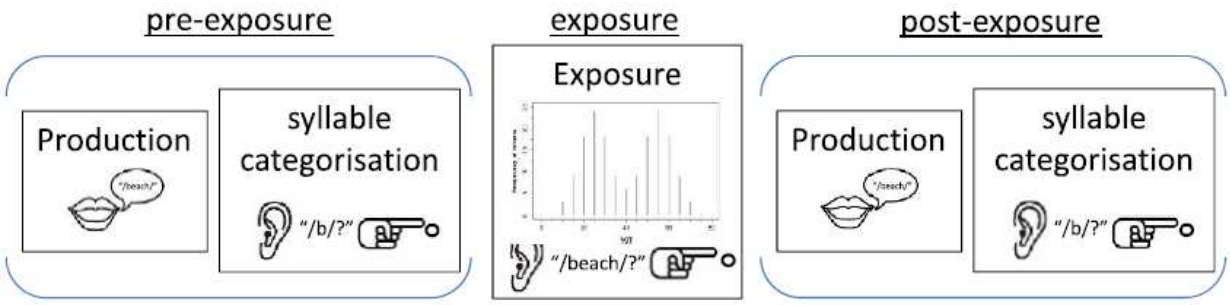


Figure 3. Left panel A: Categorization functions for distribution categorization across blocks for Experiment 1. Right panel B: Categorization functions pre- and post-exposure of the syllable categorization task in Experiment 1. Error bars show standard error.

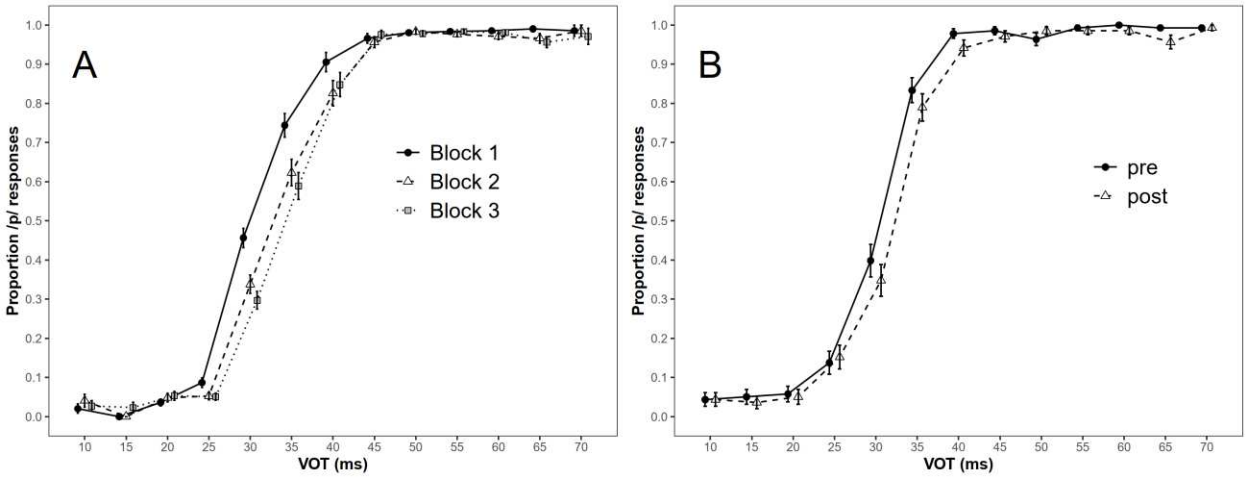


Figure 4. Top panel A: Frequency polygon for productions showing pre- and post-exposure distributions for /b/ and /p/ in Experiment 1. Bottom panel B: Empirical cumulative distribution plots for the same data. Vertical lines indicate the means for each category.

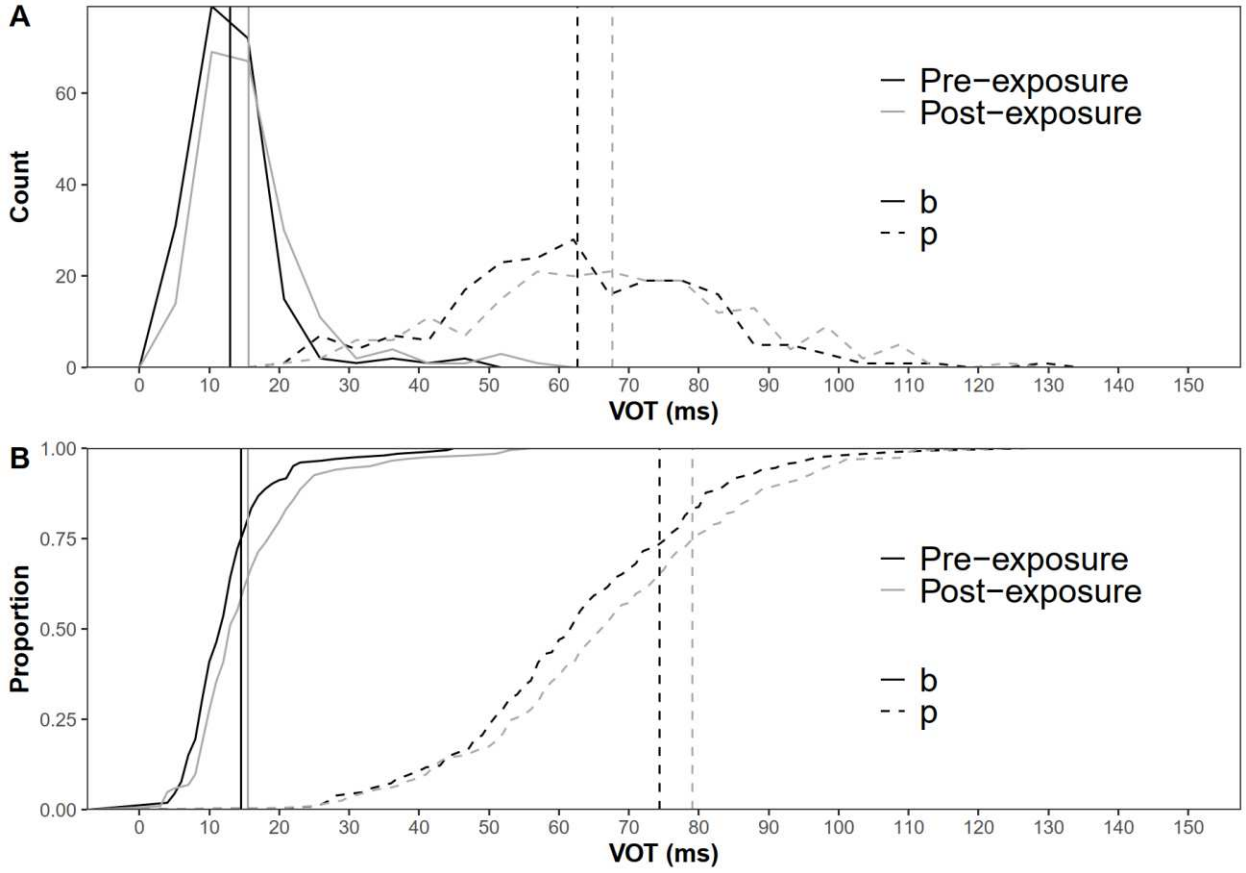


Figure 5. Left panel A: Categorization functions for distribution categorization across blocks for Experiment 2. Right panel B: Categorization functions pre- and post-exposure of the syllable categorization task in Experiment 2. Error bars show standard error.

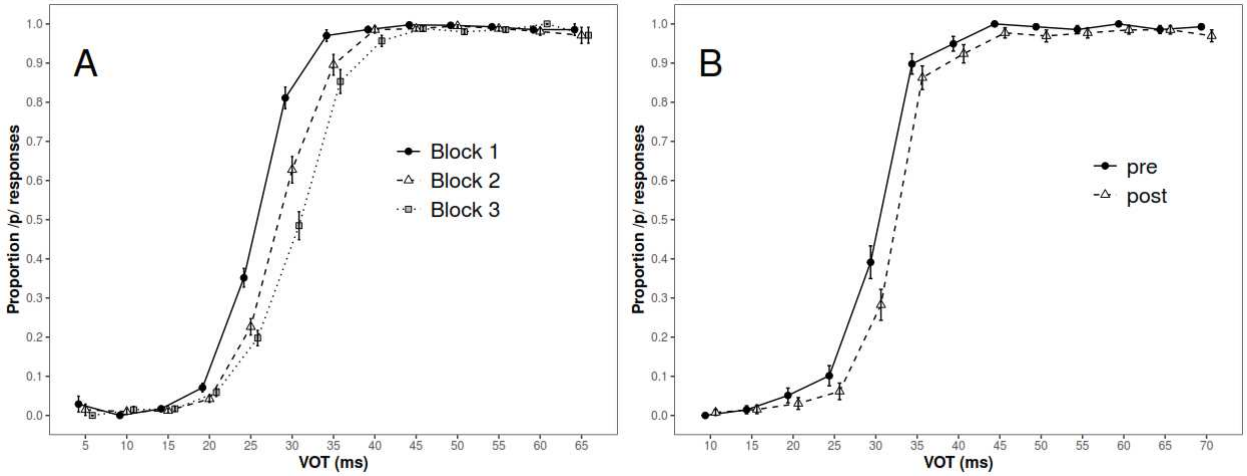


Figure 6. Top panel A: Frequency polygon for productions showing pre- and post-exposure distributions for /b/ and /p/ in Experiment 2, for both words heard and not heard in the exposure phase. Bottom panel: Empirical cumulative distribution plots for the same data. Vertical lines indicate the means for each category. Panel A (left) shows items repeated from the exposure phase, Panel B (right) shows non-repeated items.

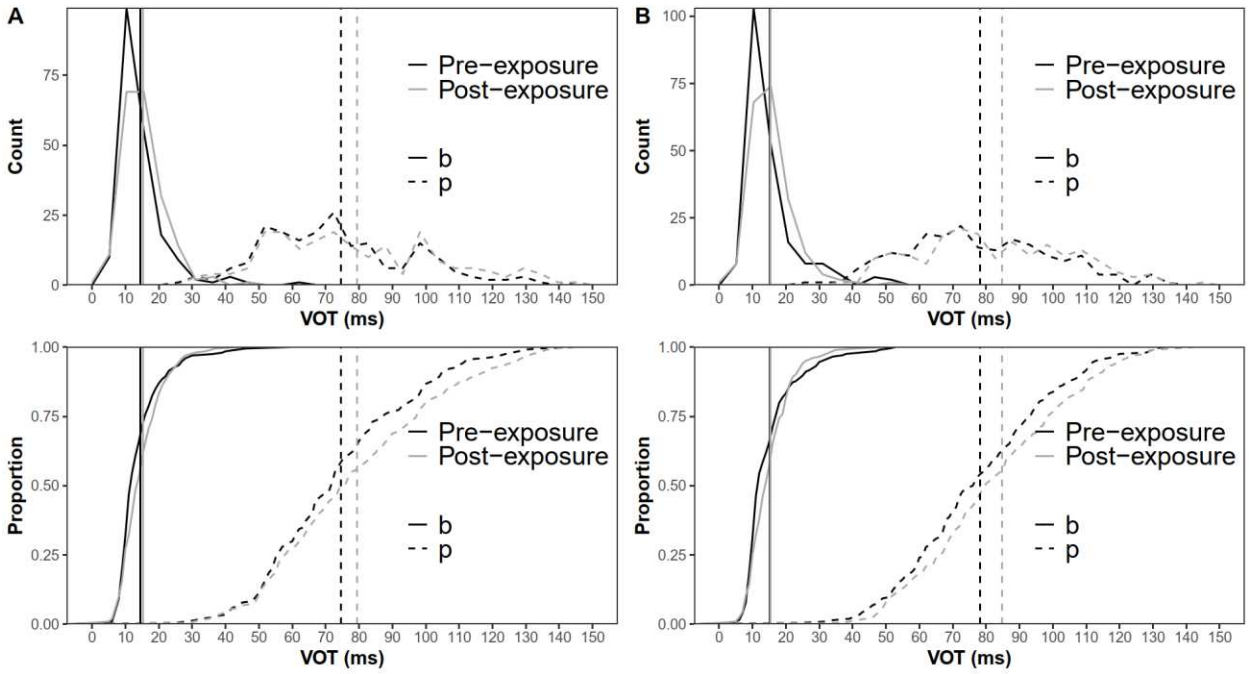


Figure 7. Histogram showing VOT distribution used in the exposure phase in Experiment 3, alongside a frequency polygon showing pre-exposure production VOT distribution from Experiment 1, scaled for comparison with our artificial distribution.

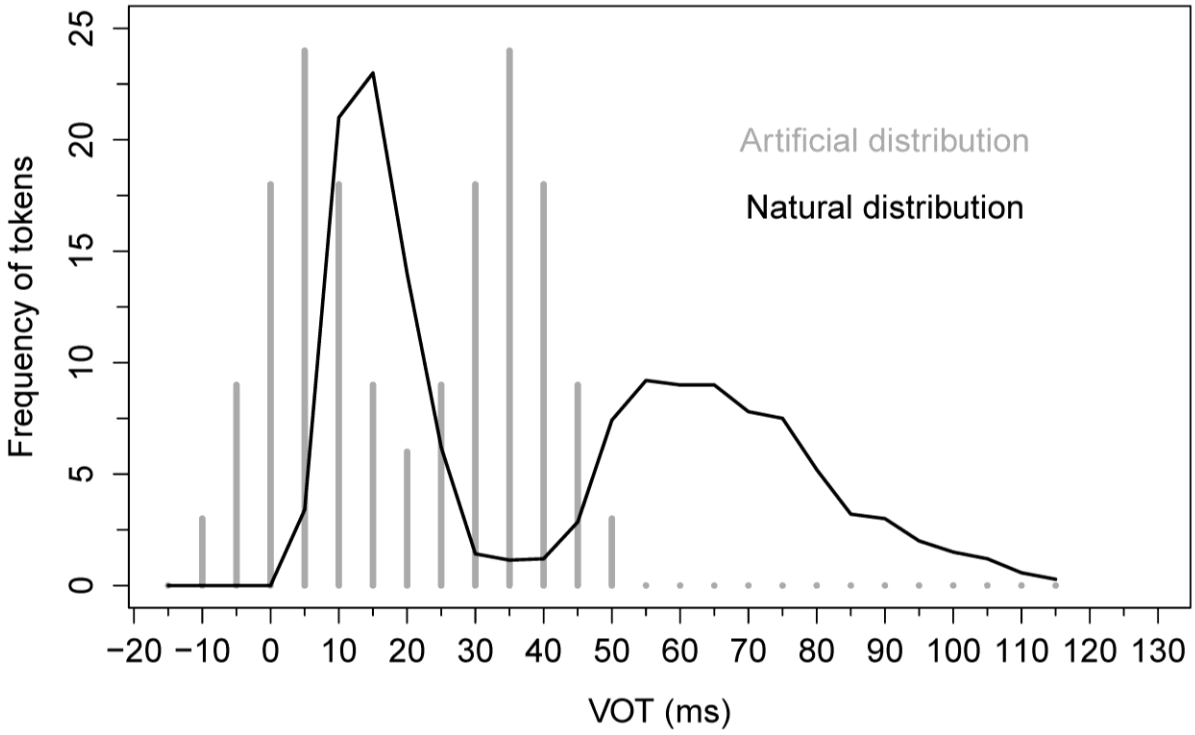


Figure 8. Left panel A: Categorization functions for distribution categorization across blocks for Experiment 3. Right panel B: Categorization functions pre- and post-exposure of the syllable categorization task in Experiment 3. Error bars show standard error.

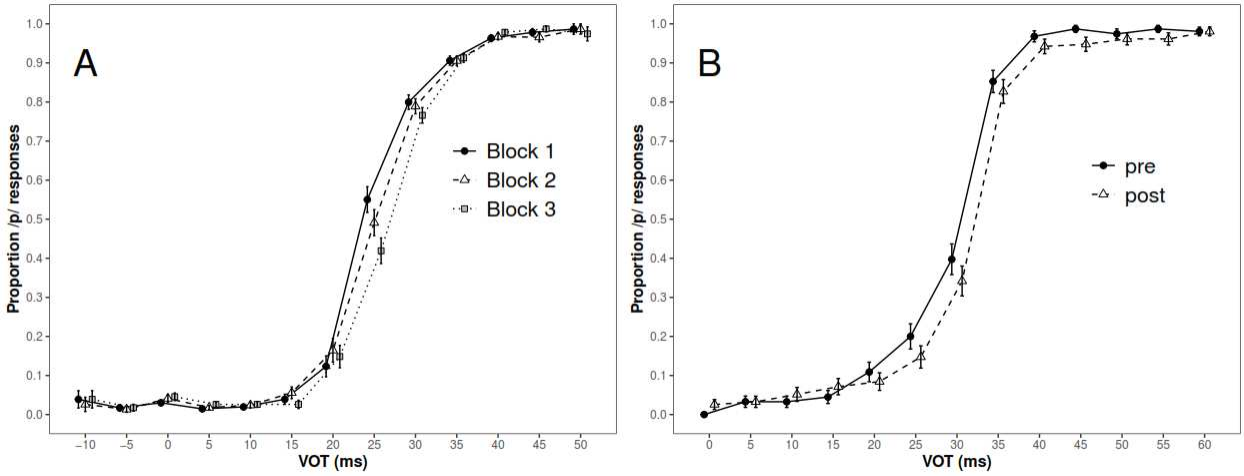


Figure 9. Top panel: Frequency polygon for productions showing pre- and post-exposure distributions for /b/ and /p/ in Experiment 3, for both words heard and not heard in the exposure phase. Bottom panel: Empirical cumulative distribution plots for the same data. Vertical lines indicate the means for each category. Panel A (left) shows items repeated from the exposure phase, Panel B (right) shows non-repeated items.

