Training non-native consonant production with perceptual and articulatory cues

Short title: Training non-native consonant production

Emily Cibelli

Department of Linguistics, Northwestern University

Acknowledgments:

This project was supported by NSF Grant DGE-1106400, and funding from the Phi Beta Kappa Northern California Association. Aaliyah Ichino, Charlotte Hoerber, Amanda Geib, Dorothy Dao, and Jocelyn Takahashi assisted with data collection and processing. The author thanks Keith Johnson, Susanne Gahl, Robert Knight, Matthew Goldrick, Jennifer Cole, Erin Gustafson, and Thomas Denby for comments on drafts of this paper. This study was conducted in the absence of any relationships that could be construed as a conflict of interest.

Address for correspondence:

Emily Cibelli, Northwestern University, 2016 Sheridan Rd., Evanston, IL, USA

Email address: emily.cibelli@northwestern.edu

Phone number: (847) 491-5831

Abstract

Background/Aims: Adult learners often struggle to produce novel phonemes in a second language, and lack clear articulatory targets. This study investigates the combined efficacy of perceptual and articulatory training, the latter involving explicit instruction about tongue position and laryngeal control, for the production of non-native phonemes.

Methods: Native English speakers were trained on a series of Hindi coronal stop consonants, with production assessed before, during, and after training sessions, on the basis of acoustic cues to place of articulation and voicing.

Results: Improvement in production was most apparent during articulatory training, when cues to target articulation were available to learners. Some improvements were maintained after training was concluded.

Conclusion: Articulatory training can contribute useful cues to pronunciation for early learners. Improvement in acquisition of targets varies in stability across learners and targets.

1 Introduction

A challenge for adults acquiring a second language is the strength of native phoneme representations, which can disrupt the acquisition of novel phonemes. Perceptually, novel categories are often assimilated to native ones (Best et al., 2001; Flege, 1995; Kuhl et al., 2008), creating a hurdle for accurate recognition and production. But even when novel categories are perceived as distinct, they may not be easy for learners to pronounce. It is not uncommon for adult second language speakers to be highly proficient in their second language, but speak with a detectable accent that reflects native language biases (Piske et al., 2001). In some cases, learners may not reliably produce a novel contrast even if they are aware that it exists.

A large body of literature is concerned with ways to improve non-native phoneme production. One approach that has received recent attention is the use of explicit production training to improve learners' metalinguistic awareness of articulatory targets. This approach reflects a view that explicit knowledge of the articulation of these sounds will allow learners to attempt to produce them with more consistent pronunciation. This study evaluates the benefit of explicit articulatory training in speakers learning to produce a set of non-native contrasts for the first time, in combination with more established perceptual training routines.

1.1 Past work on production training

While pronunciation has not always received a central focus in the L2 classroom (Fouz-González, 2015), research in second language acquisition in the past two decades has turned more attention to training paradigms aimed at improving pronunciation of non-native targets (Thomson and

Derwing, 2014). Many studies focus on perceptual training of novel contrasts, or use repetition paradigms with a native speaker; a smaller number provide explicit information to learners about gestures or articulators needed to produce a particular novel sound. Studies with classroom learners show that instruction about articulatory targets generally have beneficial effects on ameliorating L1 accents as part of the language acquisition process (Abe, 2010; Castino, 1996; Gordon et al., 2012; Lord, 2005, 2008; Saito, 2007, 2012, 2013). In the lab, visual feedback has also been explored (for a recent review, see Bliss et al., 2018). Several studies have used ultrasound as a cue to tongue position (Gick et al., 2008; Tsui, 2012; Wilson, 2014). Hazan et al. (2005) employed audiovisual training with a simulated face linked to an audio stimulus to give learners information about visually-salient articulatory postures.

In addition to studies which explicitly give feedback about articulators, several studies have given learners more abstracted information about distinct novel categories by teaching learners to interpret visually-presented acoustic displays of speech. A few studies have included training on waveform and spectrogram reading to teach learners a desired acoustic pattern (Herd et al., 2013; Saito, 2013). For vowel training, Kartushina et al. (2016) used a visual F1/F2 display to provide learners real-time feedback about vowel position and distance from native targets. Olson (2014) has demonstrated that visual training can also be used in the classroom. In this study, learners of Spanish improved their ability to produce intervocalic approximants (often mispronounced as stops by native English speakers) by learning to associate the identity and position of the segments with visual cues on a spectrogram during training.

Turning to the perceptual domain, perception training has also been used in the lab to assess links between perception and production. Bradlow et al. (1997) found that Japanese learners of English with several years of experience were able to transfer gains in perception to pronunciation, although improvement in each domain was not tightly correlated. Thomson (2011) found similar transfer at the group level for a training program of English vowels used by native Mandarin speakers. A modest improvement in production after perceptual training was also found by Baese-Berk (2010) in a study with learners being exposed to a contrast for the first time, although improvements were more substantial when production training (a repetition task with a native speaker) was also administered.

In a different vein, Catford and Pisoni (1970) took an approach to teaching novel articulations that focused on explicit information about the timing, position, and location of vocal articulators. The target contrasts were drawn from many different languages, but in all cases learners had no prior exposure to the targets. The training contained detailed explanations of the positions and movements of articulators required to produce the sounds. In their study, learners receiving this training outperformed perceptual learners on production targets, and also showed substantial gains in perceptual discrimination. This approach, which is somewhat unusual in second language acquisition but reflects the approach taken in many phonetics classrooms, may be particularly effective in raising individuals' metalinguistic awareness of the articulatory targets they are aiming to produce, as it ensures that learners know of the existence of a target, and have a concrete plan for how to produce it (whether or not they are able to execute the plan effectively).

1.2 The current study

The current study draws on a combination of approaches from past studies on pronunciation training and second language acquisition. First, it focuses on novice learners with no prior experience with the target language (e.g. Best et al., 2001), in order to measure the development of representations without prior bias or the risk of incorrect generalizations (Vlahou et al., 2012). from past exposure. Second, it concentrates on several contrasts within a single language, reflecting the work of several classroom studies (e.g. Hazan et al., 2005; Olson, 2014); this has the advantage of being more ecologically-valid than the study of a single contrast, and also taxes learners to be more precise in building representations within a tight phonetic space. Finally, similar to Catford and Pisoni (1970), it adopts an instructional approach not dissimilar to those found in phonetics classrooms and textbooks, where explicit information about articulatory postures and target gestures are given to anchor learners' attempted productions in explicit articulatory landmarks. This "phonetics textbook" approach has received relatively little attention for segmental pronunciation training in laboratory studies, although a few studies have followed up on the transfer from articulation training to perceptual discrimination (e.g. Gómez-Lacabex et al., 2008; Mathews, 1997).

By combining these components, the current study assesses novice learners' ability to acquire multiple novel contrasts in an acoustically and articulatorily dense region within a single language, using instructional training about articulatory postures in combination with perceptual training. If successful, this pronunciation training approach may prove to be a useful complement to perceptual, visual, and audiovisual training paradigms, and have utility in classrooms where the

technology for visual feedback may not always be available. To investigate this question, English speakers with no substantial experience with Hindi (or phonologically-similar languages) were recruited to learn coronal stops in Hindi.

1.2.1 Hindi and English phonology

Hindi contains a four-way voicing contrast and a dental-retroflex place of articulation contrast; English has a binary stop voicing contrast and a single (alveolar) coronal place of articulation. The Hindi stop series presents a well-documented challenge to native English speakers, particularly in identification and discrimination (Golestani and Zatorre, 2004; Pruitt et al., 2006). Pruitt et al. (2006) observed poor baseline discrimination (59%) of the Hindi dental-retroflex contrast by learners, particularly in syllable-initial position; while training improved their performance, they lagged behind native-Japanese learners in Hindi perception at both baseline and follow-up. One possible factor they cite that impedes native English learners' perception is allophonic patterns: English has phonetic-level variation in the realization of coronal stops, with more retroflexion in rhotic-adjacent stops, even though pronunciation is alveolar elsewhere. This makes it more difficult for learners to attribute dental and retroflex tokens to separate categories.

Pederson and Guion-Anderson (2010) also examined perceptual learning of place and voicing contrasts in Hindi by English speakers. They found that explicitly orienting learners' attention towards consonants (as opposed to vowels) was necessary for improvement in discrimination, lending support to the idea the explicit attention may be useful for overcoming native biases where these contrasts are concerned. However, the advantages of explicit over implicit training are not uncontroversial (Seitz et al., 2010). Vlahou et al. (2012) tested native Greek speakers on Hindi

contrasts with either explicit feedback given during a training phase, or implicit training (with or without feedback) where the consonant contrasts were not explicitly mentioned but instead paired with intensity differences. They found that learning in the implicit condition without feedback was most robust; to account for this, they suggest that explicit training and feedback may backfire in cases where learners develop incorrect generalizations about novel categories. One way to reconcile these lines of research is to posit that explicit cues must be unambiguous to learners in order for them to be used consistently and reliably.

The coronal stop inventories of Hindi and English differ in both place and voicing features. These differences make it possible to test variation in perceptual and articulatory difficulty within a single language pairing. Tees and Werker (1984) found that for English native speakers learning Hindi, perceptual discrimination of the voicing contrast was responsive to training in the short term, but improvement on the place contrast was only evident in learners with several years of exposure to the language. This suggests that different rates of acquisition may be observed over the course of the current study.

1.2.2 Predictions

Theories of non-native phoneme perception (Best et al., 2001; Flege, 1995; Kuhl et al., 2008) hypothesize that prior to learning, individuals often perceive non-native phonemes as instances of native categories, with varying degrees of fit depending on the level of mismatch between the two systems. Given this, it is expected that baseline performance, when learners have not yet been trained, will primarily reflect English biases. Table 1 catalogs the specific features of each of the coronal stops used in the current study; examples of each voicing category are shown in figure 1.

***** Table 1 about here *****

***** Figure 1 about here *****

Table 2 lays out predictions for production of each target feature before and after training, based on perceptual discrimination patterns reported in Cibelli (2015). For voicing features, the aspirated and voiceless categories are predicted to be produced accurately at pre-test, because they map onto the most-common realization of phrase-initial voiceless and voiced English stops. Voiced stops with pre-voicing appear as an allophone of English voiced (i.e. voiceless unaspirated) stops, but typically only in intervocalic position. Because participants will be producing all stops phrase-initially, it is predicted that they will not discriminate between voiced and voiceless stops at pre-test, but produce voiced stops without pre-voicing. Learning in this case would be indexed by an increase in the presence and duration of negative VOT after training.

Predictions for breathy stops are more complex, as they have pre-voicing (a feature shared with the intervocalic allophone of English /d/) and long-lag positive VOT (matching the durational properties of English /t/, but with breathy phonation). If naïve listeners are primarily sensitive to the duration of the long-lag VOT, they may perceive (and thus produce) breathy targets as instances of English /t/. If pre-voicing is most salient, they may instead produce voiceless unaspirated stops, akin to English /d/. In the latter case, an increased duration of positive VOT with a maintenance of voicing after the release burst would index learning; in both cases, increased presence and duration of negative VOT would also signal improvement.

Turning to place of articulation, it is predicted that dental and retroflex stops will not be contrasted at pre-test, because English has only a single (alveolar) coronal stop. Improvement after

training would be indexed by increased distance between the acoustic cues that signal this place contrast; in the present study, distance will be assessed using spectral properties of the stop burst and formants of the vowel following each stop.

***** Table 2 about here *****

The predictions outlined here reflect past work on L2 acquisition of Hindi showing that native English speakers often struggle to distinguish these targets from English coronal stops (Golestani and Zatorre, 2004; Pederson and Guion-Anderson, 2010; Pruitt et al., 2006; Tees and Werker, 1984). What distinguishes the current approach is a focus on acquiring the full set of contrasts within a paradigm (here, the series of Hindi coronal stops) at once, with targeted attention to multiple phonetic details. Therefore, it is worthwhile to consider the relative difficulty and learnability of each set of features. Based on past work indicating that English learners of Hindi struggle more perceptually with the dental-retroflex contrast (Werker and Tees 1984), improvement on place contrasts may be limited after perceptual training, where identification of targets must rely solely on perceptual capabilities. Performance on the place targets could accelerate during production training, when participants receive explicit instruction about tongue placement. This articulatory gesture may be more straightforward to manipulate than the laryngeal features necessary to produce the voicing contrast.

The choice to inspect the full coronal stop paradigm also permits investigation of the correlation between multiple cues at the individual level. Some learners could be more adept at producing some features than others; alternately, we may observe evidence for learners who acquire all features in tandem, and others who struggle to produce features across the paradigm.

Finally, learners may differ in the type of training that best benefits them, with some responding more to perceptual training and others to articulatory training.

2 Method

2.1 Stimuli

Consonant-vowel (CV) and vowel-consonant-vowel (VCV) syllables were recorded by a female native speaker of Hindi, with one of three vowels (/ɑ/, /i/, or /u/), selected because they are common to American English and Hindi (Ohala, 1994; Wells, 1982), and one of eight consonants (see Table 1). Two series were recorded: a "careful" series, where the speaker was instructed to speak clearly and emphasize contrasts, and a "natural" series, where the speaker was asked to recite the syllables without particular emphasis. Ten tokens of each combination of consonant, vowel, syllable structure, and style were recorded. From these 960 tokens, 384 were selected (four tokens of each combination) on the basis of an identification task conducted with two additional native speakers of Hindi, used to identify the clearest tokens. In this task, the native listeners performed an untimed eight-alternative forced-choice (8AFC) task: they heard each token and matched it to one of eight syllables written in Devanagari (Hindi orthography) that best matched their perception, with one token reflecting each of the eight coronal consonants used in the current study.

Syllables were recorded in blocks; an unintended consequence of this was that the speaker used contrastive pitch to distinguish some syllable types (e.g. / ɑd̪ʰɑ/ with low-high pitch vs. /ɑd̪ ɑ/ with high-low pitch). Instructions from the experimenter were not sufficient to eliminate this, so all final stimuli were pitch-flattened to the F0 mean across the whole stimulus set. This

consequently removed F0 correlates of voicing that are known to cue breathy stops (Hombert et al., 1979; Schiefer, 1986); however, it was considered necessary in order to avoid syllable-level pitch contours as an unintended cue to category identity. The 8AFC task described above ensured that this did not inhibit correct identification of the target consonants.

2.2 Participants

Twenty-nine native speakers of English were recruited. Ten were excluded from analysis (five for failing to complete all sessions, and five for data loss or experimenter error), leaving nineteen participants (mean age = 22.74, *s.d.* = 9.93; 15 female). Sixteen participants reported some proficiency in a second language (mean self-rated proficiency on a 4-point scale assessing reading, writing, speaking, and listening = 2.52, *s.d.* = 0.82) and nine reported some proficiency in a third language (mean = 2.50, *s.d.* = 0.70). As a group, participants reported L2 and/or L3 experience in Spanish, French, Latin, Russian, German, Mandarin, Cantonese, Chinese (variant not further specified), and American Sign Language. All participants were screened prior to enrollment to ensure that they had no proficiency in or repeated, regular experience with Hindi, whether through native fluency, classroom study, or exposure from family and friends in the home and community (regardless of whether or not they considered themselves a speaker of the language). This exclusion was extended to experience with other languages that have a four-way voicing contrast or a dental-retroflex stop contrast. Potential participants were excluded for experience with Hindi, Kannada, Marathi, Tamil, Telugu, and Urdu.

2.3 Experimental procedure

This data set was collected over eight sessions as part of a larger study designed to test perceptual and articulatory learning (perceptual results are reported in Cibelli, 2015). Table 3 summarizes the study structure. Because of evidence that sleep may aid in the consolidation of novel speech categories (Earle and Myers, 2013, 2015; Fenn et al., 2003), participants always took at least one night's break after a training session before completing a test session. The median number of days between training and testing (perception training to post-test, or production training to re-test) was 2 (range: 1-11). The median number of days to complete the full eight sessions of training and testing was 16 (range: 7-29 days). All sessions were run using custom scripts in OpenSesame (Mathôt et al., 2012). Stimuli were presented over headphones. Production responses were recorded from a stand microphone or a head-mounted condenser microphone connected to an AudioBuddy preamplifier (MAudio). Accuracy and reaction time data from the discrimination task were recorded using a serial response button box (Psychology Software Tools, Inc.).

***** Table 3 about here *****

Perception training

The four perception training sessions consisted of an AX discrimination task with trial-level accuracy feedback. These sessions were designed as a perceptual fading paradigm (Jamieson and Morosan, 1986; McCandliss et al., 2002; Protopapas and Calhoun, 2000; Terrace, 1963). This approach aims to make discrimination easy during early stages by maximizing the acoustic distance between categories, and increase the difficulty during later sessions, when cues to contrasts are more subtle. In the first training session, participants heard VCV tokens recorded in

the careful style; the second used VCV tokens in the natural style, the third used CV careful tokens, and the fourth CV natural tokens (the same stimuli used for all test sessions).

Test sessions

In test sessions (pre-test, post-test, and re-test), participants completed a repetition task to assess production performance. They listened to each of the 96 CV natural stimuli in random order, and repeated each as accurately as possible. The use of CV natural tokens, the least perceptually-distinct tokens, ensured that participants were not just benefiting from the clearest acoustic input, but adapting to the perceptual fading manipulation and becoming sensitive to tokens with less acoustic information.

Because the repetition task relies on perceptual identification to some degree – that is, participants may be better at producing targets they can recognize accurately, as the cue is auditorily-presented – performance on perception during test sessions is relevant to the interpretation of production results. While perceptual identification was not directly tested, discrimination was tested in seven of the eight sessions, giving an indication of perceptual learning. As reported in Cibelli (2015), discrimination across categories improved from pretest to post-test, and there was no change (positive or negative) in discrimination ability from post-test to re-test. This suggests that participants were better-equipped to recognize tokens in the repetition task after they completed perception training, and that they maintained this level of discrimination through the end of the experiment.

Production training

Production training gave participants explicit instruction about the articulatory gestures necessary to produce the target categories. Training was implemented as a self-paced lesson; example slides are presented in figure 2. Training began with an explanation of place of articulation: participants learned about tongue placement for the dental and retroflex consonants and how they differed from English alveolar stops. They were taught to read sagittal sections of the vocal tract, which were used along with color coding (red for retroflex, green for dental - figures 2A and 2B), as visual cues to place of articulation throughout the session.

Following this, participants were introduced to the concept of voicing, starting with the voiceless unaspirated/voiceless aspirated contrast familiar to them as the English /t/-/d/ contrast. Participants learned about the "puff of air" in aspirated consonants, and its absence during unaspirated consonants, by holding their hands in front of their face while hyperarticulating English "t" and "d". They learned visual cues for the presence and absence of aspiration (a puffing cloud and an X – see figures 2C and 2D), and practiced the distinction. Pre-voicing was introduced next, with voiced stops. Participants learned to identify the presence or absence of voicing by holding their fingers on their throat while humming, with a corresponding visual cue. When participants felt comfortable producing pre-voicing for the voiced stop, they were taught to combine pre-voicing and aspiration to produce the breathy stop. Participants then practiced combining all voicing and place of articulation features (figure 2E). At the end of the lesson, participants completed another repetition task. The task was identical to the test session repetition

task, except that the visual cue for the target consonant appeared on the screen as the participant heard the stimulus (figure 2F).

***** Figure 2 about here *****

2.4 Data processing

Syllables were annotated to define critical regions of the consonant and vowel. A first pass annotation was generated using the Penn Forced Aligner (Yuan and Liberman, 2008). Manual annotation in Praat (Boersma and Weenink, 2014) was used to correct the alignment of consonant and vowel boundaries, as the forced aligner is optimized for English targets. Sub-phonemic detail was manually annotated to mark the onset and offset of pre-voicing (when present), the onset and offset of the stop burst (when detectable), and the onset and offset of positive VOT, defined as the onset of the stop burst and the onset of periodic voicing for the following vowel, respectively. The duration of positive VOT was then calculated from the interval between these two points. When a stop burst was detected, the centroid, standard deviation, skewness, and kurtosis of the burst spectrum were extracted from the midpoint of the burst, using a custom Perl script. The IFC formant tracker (Ueda et al., 2007) was used to extract measurements of the first, second, and third formants at seven equally-spaced intervals between the vowel onset and offset.

The same acoustic properties of the stop burst and formant measurements were also extracted from the CV natural stimuli (the stimulus set used for testing and production training). This was done to provide a benchmark for interpretation of the participants' performance. The place of articulation analysis involved classification using linear discriminant analysis (LDA; see section 3.2 for detail). Performance in such a model rarely reaches 100% accuracy, even for well-

separated categories. Therefore, analysis of the stimuli, which are native speaker productions of the dental and retroflex categories, provides a basis with which to assess the maximum possible classification accuracy using the acoustic measures extracted here.

3 Results

In this study, the goal for learners is to distinguish the eight coronal consonants in the Hindi stop series; ultimately, this requires them to shift their productions away from the two alveolar stops found in English. Acoustic cues in participants' productions are used to assess whether they are achieving the target distinctions. The analysis is split by the two features that distinguish the Hindi targets from English stops: voicing and place of articulation. This approach reflects the structure of the production training paradigm, which teaches learners about each cue in sequence. Training for voicing focused primarily on duration; as such, regression models of a single continuous variable (VOT) were used to analyze the voicing data. Tongue position – the focus of place of articulation training – does not have a single most-salient acoustic correlate. Linear discriminant analysis was chosen to model place of articulation, with multiple cues (spectral moments of the stop burst; the second and third formants at vowel onset and midpoint) used in two models to predict a dichotomous outcome (dental or retroflex).

3.1 Voicing

3.1.1 Modeling approach

VOT data were analyzed from the pre-test, post-test, production training, and re-test sessions. Prior to model fitting, outliers (> 3 *s.d.* from the mean for each voicing category) were removed,

eliminating 2.33% of the data. Separate linear mixed-effects models were constructed for each voicing category (breathy, voiced, unaspirated, and unaspirated), with each of the two voicing features (aspiration/positive VOT and pre-voicing/negative VOT) as a dependent variable, resulting in eight models. All dependent variables were log-transformed.

Each model included reverse Helmert-coded fixed effects for session (comparing (1) post-test to pre-test, (2) production training to the two previous sessions, and (3) re-test to all previous sessions), contrast-coded fixed effects for place of articulation (-0.5 = dental, 0.5 = retroflex), and the interaction of place and each session predictor. Mean L2 and L3 experience (on a 4-point scale) were centered and included as control variables. Selection of the random effects structure followed Bates et al. (2015a). (Full model specifications are reported in Appendix D.) Models were fit in R (R Core Team) using the lme4 (Bates et al., 2015b) and RePsychLing (Bates et al., 2015a) packages. Each final model was re-fit after excluding extreme residuals (> 2.5 *s.d.*) (Baayen, 2008). Nested model comparisons were used to assess the significance of fixed effects.

3.1.2 Models of positive VOT

A summary of the fixed effects for the four models of positive VOT are reported in table 4. (Full model summaries for all VOT models are presented in Appendix D.) Average VOT durations at pre-test are summarized in figure 3A.

Effects of session: In the unaspirated model, positive VOT was significantly or marginally shorter at post-test (β = -0.107, $\chi^2(1) = 11.95$, $p < 0.001$), production training (β = -0.131, $\chi^2(1) = 3.73$, $p = 0.054$), and re-test (β = -0.099, $\chi^2(1) = 7.89$, $p = 0.005$). The same pattern was observed in the voiced model at post-test (β = -0.091, $\chi^2(1) = 12.72$, $p < 0.001$), production training (β = -0.102,

$\chi^2(1) = 20.86$, $p < 0.001$), and re-test ($\beta = -0.063$, $\chi^2(1) = 8.87$, $p = 0.003$). In the aspirated model, positive VOT significantly lengthened during production training only ($\beta = 0.079$, $\chi^2(1) = 18.44$, $p < 0.001$). No changes across sessions were observed in the breathy model (all $p > 0.10$); figure 3A reveals that breathy tokens were produced with long positive VOT even during the pre-test. Relative changes in each category at each session, compared to pre-test durations, are shown in figure 3B.

***** Table 4 about here *****

*****Figure 3 about here *****

Effects of control variables: Positive VOT was shorter for retroflex tokens than dental tokens for both unaspirated ($\beta = -0.152$, $\chi^2(1) = 10.49$, $p = 0.001$) and voiced ($\beta = -0.088$, $\chi^2(1) = 4.28$, $p = 0.039$) tokens. Place did not interact with session, nor were there any significant effects of L2 or L3 experience (all $p > 0.10$).

3.1.3 Models of negative VOT

Three linear mixed-effects models were constructed to assess production of negative VOT in unaspirated, voiced, and breathy stops. Only 16 of 1731 aspirated tokens had negative VOT – insufficient data to fit even a simple model. Because the dependent variable of these models contains both non-zero and zero values (i.e. tokens with either some or no pre-voicing produced), these models can be interpreted as assessing both changes in the duration of negative VOT, and the overall proportion of the presence of negative VOT.

The distribution of negative VOT was strongly bimodal, reflecting the fact that many tokens produced by participants had no pre-voicing. To test if the high proportion of zeros would skew

the negative VOT models, a two-level modeling approach was also explored. In the first level, the presence or absence of negative VOT was coded as a binary variable and analyzed using logistic mixed effects models. In the second set, linear mixed effects models were used to model the duration of negative VOT only for tokens which had a non-zero VOT value. The inferences drawn from this two-level approach were qualitatively similar to the single linear model approach; for simplicity, the single-model approach, with both zero and non-zero values in the dependent variable, is reported here. A summary of the fixed effects of these models are presented in table 5.

***** Table 5 about here *****

***** Figure 4 about here *****

Effects of session: For unaspirated tokens, there was no significant change in negative VOT across any session (all $p > 0.05$). Negative VOT lengthened during production training compared to previous sessions for voiced ($\beta = 0.924$, $\chi^2(1) = 5.88$, $p = 0.015$) and breathy ($\beta = 0.359$, $\chi^2(1) = 7.63$, $p = 0.006$) tokens. In both categories, there was a marginal or significant negative effect at re-test (voiced: $\beta = -0.331$, $\chi^2(1) = 3.55$, $p = 0.060$; breathy: $\beta = -0.241$, $\chi^2(1) = 4.17$, $p = 0.041$), indicating shorter negative VOT in the final session.

Figure 4 plots the average duration of negative VOT in voiced and breathy tokens by speaker, comparing duration of negative VOT during the pre-test to durations during subsequent test sessions. This visualization shows the average change in duration compared to baseline performance. During production training, the majority of speakers (13 of 19 for voiced; 12 of 19 for breathy) had longer average negative VOT values than during pre-test; many fewer showed this pattern at post-test (6 speakers for voiced tokens, 4 for breathy tokens). The numbers of

speakers showing lengthening drops during the re-test, as does the average duration in most individuals. However, the number of speakers above the baseline during the re-test is greater than post-test values for both categories (8 speakers for voiced tokens, 10 for breathy).

Effects of control variables: Retroflex tokens had significantly or marginally longer negative VOT for unaspirated ($\beta = 0.375$, $\chi^2(1) = 5.46$, $p = 0.019$) and voiced ($\beta = 0.289$, $\chi^2(1) = 2.78$, $p = 0.096$) tokens; there was no significant effect for breathy tokens ($p > 0.05$). The effects of L2 and L3 experience, and all interactions, failed to reach significance in all models (all $p > 0.05$), with the exception of the interaction of place and session (pre-test vs. post-test) in the breathy model ($\beta = 0.368$, $\chi^2 = 5.36$, $p = 0.021$). This was driven by longer pre-voicing for dental tokens at pre-test (mean dental negative VOT: 669 ms; mean retroflex: 583 ms), but longer pre-voicing for retroflex tokens in the post-test (mean dental: 403 ms; mean retroflex: 654 ms).

3.2 Place of articulation

To assess production of the dental/retroflex contrast, two sets of acoustic features were extracted: formant frequencies of the vowel following the stop, and spectral properties of the stop burst. For each feature set, linear discriminant analysis (LDA) with leave-one-out cross-validation was used to assess the separability of the dental and retroflex categories at each session. Permutation tests with 1000 repetitions were used to identify the likelihood that a particular accuracy value would be achieved by chance (Combrisson and Jerbi, 2015).

3.2.1 Formants

Formants at vowel onset provide cues to the place of the preceding consonant (Delattre et al., 1955; Kewley-Port, 1982; Liberman et al., 1954), and formants at vowel midpoint may also hold

information about consonant identity (Sussman et al., 1991, 1993). Because retroflexion lowers F3 (Stevens and Blumstein, 1975; Werker et al., 1981; Werker and Tees, 1984), F2 and F3 at vowel onset and midpoint were considered in the current analysis. Formant analyses were restricted to voiced and voiceless unaspirated tokens, as long positive VOT following the burst can obscure the relationship between consonant place and formants.

An LDA model of the CV unaspirated stimuli was constructed to provide a benchmark for classification accuracy in the speech of an native speaker. A preliminary generalized linear model was used to identify significant predictors out of the set of F2 onset, F2 midpoint, F3 onset, and F3 midpoint, with place (dental or retroflex) as the outcome variable. F2 onset and F3 midpoint were the only significant predictors; using these as features in the LDA model, 71.4% of stimuli were correctly classified.

In the participant data, outliers (tokens with F2 or F3 values > 3 *s.d.* of a vowel category mean) were removed. The same feature selection procedure was then applied. Separate predictors were selected for participant data because it is possible that learners may use a different combination of acoustic cues than the model native speaker to signal the dental/retroflex contrast; this liberal approach provides the best chance for learners to demonstrate a contrast. The final model included F2 onset and F3 onset as predictors. These were used as features in four LDA analyses, one for each session.

A series of 1000 permutation tests were run for each session to assess the likelihood of observing the accuracy values by chance. In each test, the link between formant data and category label were scrambled, and the classifier re-fit to this permuted data. Performance was assessed as

the number of tests where the classifier of the true data was more accurate than the classifier of the permuted data. If the permutations met or exceeded the classifier no more than 5% of the time, the LDA model was considered to be significantly different from chance for that session. LDA models were also constructed by-participant, to assess the variability in the fit of these features to any one individual's data. The distributions of classification accuracy for the by-participant models are shown in figure 5A.

***** Table 6 about here *****

Accuracy and test results are reported in Table 6. There was no session in which speakers reliably distinguished dental from retroflex consonants on the basis of the formants of the following vowel. Results from the production training session were marginal (only 6.1% of random permutations beating the classifier), with 59.1% of tokens correctly classified. However, at re-test the classification accuracy dropped down to 52.8% and did not differ from chance.

To compare performance across sessions, the decision of the classifier for each token was compared to the token's true identity, to generate an accuracy code (correct or incorrect). This was used as the dependent variable in a logistic mixed-effects model. The model included three Helmert-coded fixed effects for session, comparing the (1) post-test, (2) production training, and (3) re-test sessions to all previous sessions. It also included by-subject random slopes for the production training session effect. The model found no change in accuracy at post-test (β = -0.044, $\chi^2(1) = 0.21$, $p > 0.010$). There was a significant increase in accuracy in production training (β = 0.358, $\chi^2(1) = 0.36$, $p = 0.002$) and a marginal decrease in accuracy at re-test (β = -0.128, $\chi^2(1) = 2.77$, $p = 0.096$), replicating the pattern of classification accuracies.

*****Figure 5 about here *****

3.2.2 Burst spectra

The dental-retroflex contrast is also reliably cued by spectral properties of the stop burst (Blumstein and Stevens, 1979; Kewley-Port et al., 1983). These cues have the advantage of not being disrupted by long-lag positive VOT, allowing the entire data set to be investigated. Following the analysis in Forrest et al. (1988), four spectral moments were measured from a spectrum extracted at center of each stop burst: centroid, standard deviation (variance), skewness, and kurtosis. The acoustics of the experiment stimuli were again measured to provide a native-speaker benchmark for comparison to participants' speech. All stimuli where a burst longer than 2 ms could be identified (338 of 384 stimuli) were entered into a generalized linear model predicting place of articulation, with the four linear spectral moments as predictors. The model with all four moments was the best fit to the stimulus data. Using these features as the input to an LDA model yielded a classification accuracy of 75.1%.

The same procedure was used to assess the participant data, again after outlier removal (tokens with values > 3 *s.d.* for any measure). A generalized linear model fit indicated that only standard deviation and skewness were reliable predictors of place. These two features were entered into separate cross-validated LDA models for each session, with permutation tests to assess whether accuracy differed from chance. Results are reported in table 7. Each participant's data were also classified individually using the same features; individual performance by session is plotted in figure 5B.

***** Table 7 about here *****

Classification accuracy was above chance in both the pre-test and post-test sessions (53.9% and 55.1% accuracy, respectively), indicating that even at first exposure, participants were making some distinction between dental and retroflex tokens. Accuracy reached 65% during the production training session, but dropped below pre-test levels during the re-test (53.4%, not significantly different from chance).

As with the formant analysis, a mixed-effects logistic model was used to compare accuracy across sessions. There was no significant change in accuracy at post-test ($\beta = 0.041$, $\chi^2(1) = 0.34$, $p = 0.557$). Accuracy was significantly better during production training ($\beta = 0.460$, $\chi^2(1) = 14.65$, $p < 0.001$) but significantly decreased at re-test ($\beta = -0.195$, $\chi^2(1) = 11.22$, $p < 0.001$).

3.3 Individual variation in feature production

The trends reported above indicate how participants performed on individual features, but performance *across* features may not be consistent across individuals. It is known that English-speaking learners of Hindi especially struggle with perceiving the place contrast compared to the voicing contrast. Tees and Werker (1984) found that short-term laboratory training with English speakers with no Hindi experience was sufficient to improve discrimination of a Hindi voice contrast, but not a place contrast. They report a similar difficulty for place contrasts for speakers with one or two years of Hindi study (although learners with five years of experience showed improvement in both contrasts). Because of this pattern, and because production training in the current study took distinct approaches to teaching each cue, we might expect to see different performance in the production of each type of contrast (voicing vs. place). The place contrast relied on visual cues – a sagittal section – to teach learners about tongue position. For the voicing

contrasts, the hand was used as a tactile cue to voicing and aspiration. There is something of a trade-off in cue difficulty for learners here – the place contrast is more difficult to hear, but may be more intuitive to produce as a novice. Because of these differences, the success rate in learning each class may vary. Furthermore, participants may vary in which type of training is most beneficial or intuitive to them.

To compare performance on place and voicing features across individuals, a binary metric of success was created for each feature in each token. For place features (burst spectra and formants), the by-token classification accuracy from each LDA model was used. For voicing features – pre-voicing in voiced and breathy tokens, and positive VOT in breathy tokens – a threshold of successful production was established. For pre-voicing, any token with non-zero negative VOT was considered to be successful. For positive VOT in breathy tokens, VOTs greater than 30 ms were considered successful; this threshold was chosen to ensure that tokens were distinct from short-lag (unaspirated) tokens.

To compare links between feature performance across sessions, correlations were run for each feature at each session. For each feature and session combination, one data point represented the proportion of successful tokens of that feature produced by one participant. Correlations are plotted in figure 6.

*****Figure 6 about here *****

Several correlations emerged between voicing features. Production of negative VOT in voiced and breathy tokens was positively correlated at pre-test ($r = 0.755$, *adj.* $p < 0.001$, FDR correction applied to all correlations) and maintained at post-test ($r = 0.599$, *adj.* $p = 0.013$),

production training ($r = 0.653$, *adj.* $p = 0.005$), and re-test ($r = 0.505$, *adj.* $p = 0.050$). The presence of negative VOT in breathy and voiced tokens was negatively correlated with positive VOT in breathy tokens; this trend was consistent across sessions, and significant in all but one case (re-test, breathy positive VOT and voiced negative VOT, $r = -0.393$, *adj.* $p = 0.159$). Participants who were more likely to correctly produce pre-voicing were *less* likely to produce long-lag positive VOT on breathy tokens. Turning to place features, significant or marginal correlations only emerged during production training, when correct formant classification was positively correlated with voiced negative VOT ($r = 0.481$, *adj.* $p = 0.065$) and burst classification ($r = -0.612$, *adj.* $p = 0.011$), but negatively correlated with breathy positive VOT ($r = -0.507$, *adj.* $p = 0.050$).

4 Discussion

This study tested the combined effects of two types of training on the production of a non-native series of stop consonants by adult learners. Of interest was the impact of a methodology that included both perceptual and articulatory training, as well as differences between the acquisition of different articulatory features. It was predicted that at baseline, learners would assimilate non-native categories to native categories. After training, listeners were predicted to increase the presence of negative VOT in breathy and voiced tokens, and the duration of positive VOT in breathy tokens, to distinguish these categories from the unaspirated and aspirated voiceless categories present in their native language. Learners were also predicted to assimilate dental and retroflex tokens to a single coronal category at pre-test, and to make a distinction between them

after training. Analyses of individuals were used to assess whether there were by-participant relationships in performance across feature categories.

4.1 Findings

4.1.1 Summary of voicing results

In the analysis of negative VOT, the prediction that participants would lengthen pre-voicing in breathy and voiced tokens was borne out primarily in the production training session, with retention into re-test for some speakers. The positive VOT analysis showed that for breathy targets, learners produced long positive VOTs at baseline and did not change throughout the study, indicating that they were sensitive to long-lag aspect of breathy stops even at pre-test.

Interestingly, the production of positive VOT in native categories (unaspirated and aspirated voiceless stops) showed an enhancement the long-lag/short-lag voicing contrast over the course of training. In other words, there was a larger duration difference for the two voicing categories found in English. This finding was not predicted, but may reflect an increase in the overall precision of VOT targets in response to the increased complexity of the novel four-way paradigm.

4.1.2 Summary of place of articulation results

Dental and retroflex tokens were accurately classified significantly above chance during production training. However, this benefit was not maintained at the re-test session for many speakers, and the group as a whole. This pattern provides evidence for the efficacy of training for the novel place of articulation contrast, but does not establish that it will leave a lasting effect on production targets. Across the two sets of cues, the burst spectra analysis showed more sensitivity

to the contrast than the formant analysis, with significant (if modest) classification in the pre- and post-test sessions as well.

4.1.3 Links between place and voicing performance at the individual level

Strong positive correlations between negative VOT in breathy and voiced tokens indicate that participants who were successful at producing pre-voicing in one category were consistent in applying it to the other. A *negative* correlation between negative VOT and positive VOT in breathy targets revealed that participants who were more likely to correctly produce pre-voicing were *less* likely to produce long-lag positive VOT. Put another way, participants were not coupling long-lag positive VOT with pre-voicing, suggesting that their breathy productions were like either aspirated [t^h] or plain voiced [d]. This indicates a split production of the breathy stop: participants latched onto one feature or the other, but were unlikely to unite them.

Correlations with place features were more variable, and tended to be reliable only during production training. In that session, participants who were performing well on production of pre-voicing were also able to distinguish the two place categories. This suggests that those who produced pre-voicing reliably were more successful at the task overall than those who were focused solely on long-lag positive VOT in breathy tokens, as the latter was negatively correlated with both place and other voicing features.

4.1.4 Key takeaways

Taken together, these findings provide evidence that participants changed their production of novel targets during training, although the persistence of this effect into the re-test session was mixed. Improvements were most apparent during production training, and more subtle at the post-test

session (after perception training). However, it is possible that perception training played a key role, but that effects were not evident by the time of the post-test. All improvement during production training occurred after perception training, so the combined effects of both training types may be responsible for performance in later sessions. And while the explicit cues to articulation during production training might provide the greatest support to learners, there is evidence in the literature to support the idea that transfer from perception training also played a role. Bradlow et al. (1997) found a link from perception training to production performance at the group level. And while these speakers had more experience with their L2 than in the current study, Baese-Berk (2010) also found a transfer from perception training to production for a novel VOT contrast in learners without prior experience. These findings suggest that the speakers in the current study may have also received benefit from perception training, but the experimental paradigm did not permit explicit comparisons of improvement after each training type individually.

Despite individual variation in the production of novel features, a general pattern emerged. In production of the breathy category, participants tended to either produce negative VOT or long-lag positive VOT – but not both – suggesting that they were attuned to either pre-voicing or aspiration. One possibility to explain this may be an initial perceptual bias to perceive this category as either English /d/ (which has a pre-voiced allophone in medial position) or English /t/ (which has long-lag positive VOT). The presence of pre-voicing was positively correlated with formant cues to place of articulation during production training; given that, it may be said that the group of individuals producing pre-voicing were "more successful" across all novel features. These patterns were generally apparent even at pre-test, suggesting that strong perceptual skills may naturally

give an advantage to some learners. However, almost all features and learners showed a trend towards improvement during production training, indicating a role for articulatory instruction regardless of initial (dis)advantages.

Improvements in pronunciation were generally followed by decreased performance during the re-test session, although there was substantial individual variation. In the place of articulation analysis, the dental-retroflex contrast returned to chance levels when looking at both formant and burst features. In the analysis of negative VOT, several speakers did retain some production of pre-voicing for breathy and voiced tokens, but the mean duration was reduced compared to previous sessions. One factor impacting performance during the re-test is the relative instability of perceptual representations. While production training provided visual cues, during the re-test (and other testing sessions) listeners had to rely on acoustic cues from the stimulus to identify the target. This may be particularly challenging for the place of articulation contrast (Tees and Werker, 1984). And while learners in this population did not show a drop in their perceptual performance from post-test to re-test (reported in Cibelli 2015), they also did not perform at ceiling; therefore, their perceptual representations may not have been stable enough to support formulation of an accurate articulatory plan. Furthermore, because participants were only tested on discrimination and not identification, it is possible that they improved their ability to discriminate without a concurrent improvement in their ability to identify the *precise* category of a target. In other words, an ability to detect a distinction between two tokens does not necessarily mean that learners were certain of *which* categories they were perceiving – or that they could accurately produce them. This suggests

an advantage for of articulatory training, as it circumvents instability in perceptual representations and gives learners direct information about production targets.

4.2 Considerations for future study

Several questions are left open by the current findings. One natural direction to pursue would be a direct comparison of perceptual and articulatory training. The current study integrates both methods into a single training paradigm, making it impossible to directly compare the efficacy of each. A between-subjects study that exposed learners to only one approach could clarify the relative contribution of each to the early acquisition of non-native targets.

A limiting factor in the current analysis is the reliance on acoustic, rather than perceptual, assessment of participants' productions. While the acoustics measured reflect cues thought to be crucial to the target contrasts, it is possible that the measured changes do not reflect a linear movement towards more native-like production. Thomson and Derwing (2014) note that changes by non-native speakers after training may reflect increased native-like production, greater intelligibility to native speakers, or simply greater acoustic discriminability; it is unclear which of these applies to the current results. Conversely, some cues may be present that are not well-represented by the current analyses, but which reflect a more discriminable production of these contrasts. Future work using perceptual judgments by native Hindi speakers could clarify whether production is becoming more "Hindi-like," or more perceptually discriminable, and not simply more acoustically contrastive.

These findings highlight substantial individual variation in the acquisition and retention of certain features. While experience with a second or third language did not predict performance,

there may be other individual-level factors that underlie this variation. For example, it is possible that some individuals have more awareness of the position and control of their articulators than others; if so, this undoubtedly affects how well they are able to implement the cues presented in this approach. Other types of interventions focused on articulatory learning, such as live imitation of a native speaker, or visual feedback with tools such as ultrasound imaging (Gick et al., 2008; Tsui, 2012; Wilson, 2014) may be more suitable for some learners. In addition, divergent performance on the non-native features of the breathy token may suggest that early (mis)perceptions of non-native tokens vary by individual. Given that, strategies (both perceptual and articulatory) that identify individual challenges at the beginning of training and draw a learner's attention to them may be more efficient.

Finally, some limitations in participants' performance may originate in incomplete perceptual representations. For learners who do not have stable perceptual categories, their performance in any session where visual cues are not present will be limited by their ability to use acoustic cues to identify the target. However, learners with more experience in the language may have stronger perceptual representations, making them more ideal candidates to show improvement in repetition tests where explicit cues are not present. A replication of this paradigm with intermediate or highly proficient L2 speakers who retain L1 accents could test this prediction.

5 Conclusions

This study supports the claim that explicit articulatory training can be effectively integrated into a perceptual training paradigm for the acquisition of novel production targets in a second language.

This approach may be an efficient way to jumpstart learning for speakers who are new to a contrast and who do not yet have stable perceptual categories. However, because improvements were not retained by all speakers once visual cues were no longer present, future work is needed to reveal the conditions under which this type of training will lead to stable improvements in the production of novel phonemes.

References

Abe H (2010): Form-focused instruction in L2 pronunciation pedagogy: The effect of negotiation of form in a Japanese classroom. In *Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech New Sounds. Adam Mickiewicz University*, pp 1–6.

Baayen R H (2008): *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baese-Berk M M (2010): *An examination of the relationship between speech perception and production*. PhD thesis, Northwestern University.

Bates D, Kliegl R, Vasishth S, Baayen H (2015a): Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates D, Mächler M, Bolker B, Walker, S (2015b): Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Best CT, McRoberts GW, Goodell E (2001): Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America* 109(2):775–794.

Bliss H, Abel J, Gick B (2018): Computer-assisted visual articulation feedback in L2 pronunciation instruction. *Journal of Second Language Pronunciation* 4(1):129-153.

Blumstein SE, Stevens KN (1979): Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America* 66(4):1001–1017.

Boersma P, Weenink, D (2014): Praat: Doing phonetics by computer (version 5.1.13).

Bradlow AR, Pisoni DB, Akahane-Yamada R, Tohkura Y (1997): Training Japanese listeners to identify English /r/ and /l/ IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America* 101(4):2299–2310.

Castino J (1996): Impact of a phonetics course on FL learners' acquisition of Spanish phonology. *Selecta: Journal of the Pacific Northwest Council on Foreign Languages* 17:55–58.

Catford JC, Pisoni DB (1970): Auditory vs. articulatory training in exotic sounds. *The Modern Language Journal* 54(7):477–481.

Cibelli E (2015): *Aspects of articulatory and perceptual learning in novel phoneme acquisition*. PhD thesis, University of California, Berkeley.

Combrisson E, Jebri K (2015): Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods* 250:126–136.

Delattre PC, Liberman AM, Cooper FS (1955): Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America* 27(4):769–773.

Earle FS, Myers EB (2013): Building phonetic categories: an argument for the role of sleep. *Frontiers in Psychology* 5:1192.

Earle FS, Myers EB (2015): Sleep and native language interference affect nonnative speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6):1680–1695.

Fenn KM, Nusbaum HC, Margoliash D (2003): Consolidation during sleep of perceptual learning of spoken language. *Nature* 425(6958):614–616.

Flege JE (1995): Second language speech learning: Theory, findings, and problems. In Strange W (eds): *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, York Press, Timonium, pp 233–277.

Forrest K, Weismer G, Milenkovic P, Dougall RN (1988): Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America* 84(1):115–123.

Fouz-González J (2015): Trends and directions in computer-assisted pronunciation training; in Mompean J, Fouz-González J (eds), *Investigating English Pronunciation*. London, Palgrave Macmillan, pp 314-342.

Gick B, Bernhardt B, Bacsfalvi P, Wilson I, Hansen Edwards J, Zampini, M (2008): Ultrasound imaging applications in second language acquisition. *Phonology and Second Language Acquisition* 36:315–328.

Golestani N, Zatorre RJ (2004): Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage* 21(2):494–506.

Gómez-Lacabex EG, García Lecumberri M, Cooke M (2008) Identification of the contrast full vowel- schwa: training effects and generalization to a new perceptual context. *Ilha do Desterro* 55: 173–196.

Gordon J, Darcy I, Ewert, D (2012): Pronunciation teaching and learning: Effects of explicit phonetic instruction in the L2 classroom; in Levis J, LeVelle K (eds): *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*. Ames, Iowa State, pp 194–206.

Hazan V, Sennema A, Iba M, Faulkner A (2005): Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47(3):360–378.

Herd W, Jongman A, Sereno J (2013): Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. *The Journal of the Acoustical Society of America* 133(6):4247–4255.

Hombert JM, Ohala JJ, Ewan WG (1979): Phonetic explanations for the development of tones. *Language* 55(1):37–58.

Jamieson DG, Morosan DE (1986): Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by Francophones. *Perception & Psychophysics* 40(4):205–215.

Kartushina N, Hervais-Adelman A, Frauenfelder UH, Golestani N (2016): Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics* 57:21–39.

Kewley-Port D (1982): Measurement of formant transitions in naturally produced stop consonant–vowel syllables. *The Journal of the Acoustical Society of America* 72(2):379– 389.

Kewley-Port D, Pisoni DB, Studdert-Kennedy M (1983): Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America* 73(5):1779–1793.

Kuhl PK, Conboy B, Coffey-Corina S (2008): Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e): *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:979–1000.

Liberman AM, Delattre PC, Cooper FS, Gerstman LJ (1954): The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied* 68(8):1–13.

Lord G (2005): (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania* 88(3):557–567.

Lord G (2008): Podcasting communities and second language pronunciation. *Foreign Language Annals* 41(2):364–379.

Matthews J (1997): The influence of pronunciation training on the perception of second language contrasts. *International review of Applied Linguistics*, *35*(2):223-229.

Mathôt S, Schreij D, Theeuwes J (2012): OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2):314–324.

McCandliss BD, Fiez JA, Protopapas A, Conway M, McClelland JL (2002): Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience* 2(2):89–108.

Ohala M (1994): Hindi. *Journal of the International Phonetic Association*. *24*(1): 35-38.

Olson DJ (2014): Phonetics and technology in the classroom: a practical approach to using speech analysis software in second-language pronunciation instruction. *Hispania* 97(1):47–68.

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

Piske T, MacKay IR, Flege JE (2001): Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics* 29(2):191–215.

Protopapas A, Calhoun B (2000): Adaptive phonetic training for second language learners. In *Proceedings of the 2nd International Workshop on Integrating Speech Technology in Language Learning*, pages 31–38.

Pruitt JS, Jenkins J, Strange W (2006): Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America* 119(3):1684–1696.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saito K (2007): The influence of explicit phonetic instruction on pronunciation in EFL settings: The case of English vowels and Japanese learners of English. *Linguistics Journal* 3(3):16–40.

Saito K (2012): Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly* 46(4):842–854.

Saito K (2013): Reexamining effects of form-focused instruction on L2 pronunciation development. *Studies in Second Language Acquisition* 35(1):1–29.

Schiefer L (1986): F0 in the production and perception of breathy stops: Evidence from Hindi. *Phonetica* 43(1-3):43–69.

Seitz AR, Protopapas A, Tsushima Y, Vlahou EL, Gori S, Grossberg S, Watanabe T (2010): Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition* 115(3):435–43.

Stevens KN, Blumstein SE (1975): Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics* 3:215–233.

Sussman HM, Hoemeke KA, Ahmed FS (1993): A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *The Journal of the Acoustical Society of America* 94(3):1256–1268.

Sussman HM, McCaffrey HA, Matthews SA (1991): An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America* 90(3):1309–1325.

Tees RC, Werker JF (1984): Perceptual flexibility: maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie* 38(4):579.

Terrace HS (1963): Discrimination learning with and without "errors". *Journal of the Experimental Analysis of Behavior* 6(1):1–27.

Thomson RI (2011): Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal* 28(3):744–765.

Thomson RI, Derwing TM (2014): The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics* 36:326–334.

Tsui HML (2012): Ultrasound speech training for Japanese adults learning English as a second language. Master's thesis, University of British Columbia.

Ueda Y, Hamakawa T, Sakata T, Hario S, Watanabe A (2007): A real-time formant tracker based on the inverse filter control method. *Acoustical Science and Technology* 28(4):271–274.

Vlahou EL, Protopapas A, Seitz AR (2012): Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General* 141(2):363–381.

Wells, JC (1982): Accents of English, Volume 3: Beyond the British Isles. Cambridge, 467-674.

Werker JF, Gilbert JH, Humphrey K, Tees RC (1981): Developmental aspects of cross-language speech perception. *Child Development* 52:349–355.

Werker JF, Tees RC (1984): Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1):49–63.

Wilson I (2014): Using ultrasound for teaching and researching articulation. *Acoustical Science and Technology* 35(6):285–289.

Yuan J, Liberman M (2008): Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*.

Appendix A Table headings, main sections

Table 1: The eight Hindi coronal stop consonants.

Table 2: Predictions for performance on each target feature (voicing and place of articulation) before training (pre-test) and after training.

Table 3: Structure of the experiment.

Table 4: Summary of effects for positive VOT models. The direction of significant effects ($p < 0.05$, as assessed with likelihood ratio tests of models with and without that predictor) are reported in plain text; marginal effects ($p < 0.10$) are italicized. Full model estimates are reported in Appendix D.

Table 5: Summary of effects for negative VOT models. Significant effects ($p < 0.05$, as assessed with likelihood ratio tests) are reported in plain text, marginal effects ($p < 0.10$) are italicized. Only the direction of effects is noted; full model estimates are reported in Appendix D. Note that there was an insufficient number of non-zero data points to construct a model for aspirated tokens.

Table 6: Classification results and results of permutation tests by session, formant data. Sessions where no more than 5% of the permutation tests meet or exceed the classifier accuracy are considered to be significantly different from chance; 10% represents a marginal threshold.

Table 7: Classification results and results of permutation tests by session, formant data. Sessions where no more than 5% of the permutation tests meet or exceed the classifier accuracy are considered to be significantly different from chance; 10% represents a marginal threshold.

Appendix B Tables, main sections

Consonant	Voicing	Positive VOT	Negative VOT	Place
t̪	unaspirated	short lag	none	dental
t̪ʰ	aspirated	long lag	none	dental
d̪	voiced	short lag	pre-voicing	dental
d̪ʱ	breathy	long lag (breathy)	pre-voicing	dental
ʈ	unaspirated	short lag	none	retroflex
ʈʰ	aspirated	long lag	none	retroflex
ɖ	voiced	short lag	pre-voicing	retroflex
ɖʱ	breathy	long lag (breathy)	pre-voicing	retroflex

Table 1: The eight Hindi coronal stop consonants.

Feature	Prediction at baseline	Prediction after training
Aspirated	Accurate production (comparable to English /t/)	No change
Voiceless	Accurate production (comparable to word-initial English /d/)	No change
Voiced	Assimilate to voiceless (allophone of English /d/)	Increased negative VOT duration
Breathy	Assimilation to either English /t/ or /d/	Increased negative VOT duration; if assimilation to /d/ at pre-test, increased positive VOT duration
Dental	Assimilation to single category with retroflex	Distinguishable formant/burst acoustic cues from retroflex
Retroflex	Assimilation to single category with dental	Distinguishable formant/burst acoustic cues from dental

Table 2: Predictions for performance on each target feature (voicing and place of articulation) before training (pre-test) and after training.

Session	Perception task	Perception feedback	Production task	Production feedback	Stimuli
Pre-test	AX discrimination	-	Repetition	-	CV natural
Perception training 1	AX discrimination	Accuracy feedback	-	-	VCV careful
Perception training 2	AX discrimination	Accuracy feedback	-	-	VCV natural
Perception training 3	AX discrimination	Accuracy feedback	-	-	CV careful
Perception training 4	AX discrimination	Accuracy feedback	-	-	CV natural
Post-test	AX discrimination	-	Repetition	-	CV natural
Production training	-	-	Repetition	Visual cues	CV natural
Re-test	AX discrimination	-	Repetition	-	CV natural

Table 3: Structure of the experiment. Production data reported in the current study is drawn from the four sessions that contained a repetition production task.

	Unaspirated	Aspirated	Voiced	Breathy
Mean L2 experience	n.s.	n.s.	n.s.	n.s.
Mean L3 experience	n.s.	n.s.	n.s.	n.s.
Session: pre-test vs. post-test	negative	n.s.	negative	n.s.
Session: pre/post-test vs. prod training	*negative*	positive	negative	n.s.
Session: all previous vs. re-test	negative	n.s.	negative	n.s.
Place of articulation (POA)	negative	n.s.	negative	n.s.
POA*session (pre-test vs. post-test)	n.s.	n.s.	n.s.	n.s.
POA*session (pre/post-test vs. prod training)	n.s.	n.s.	n.s.	n.s.
POA*session (all previous vs. re-test)	n.s.	n.s.	n.s.	n.s.

Table 4: Summary of effects for positive VOT models. The direction of significant effects ($p < 0.05$, as assessed with likelihood ratio tests of models with and without that predictor) are

reported in plain text; marginal effects ($p < 0.10$) are italicized. Full model estimates are reported in Appendix D.

	Unaspirated	**Aspirated**	**Voiced**	**Breathy**
Mean L2 experience	n.s.	-	n.s.	n.s.
Mean L3 experience	n.s.	-	n.s.	n.s.
Session: pre-test vs. post-test	n.s.	-	n.s.	n.s.
Session: pre/post-test vs. prod. training	n.s.	-	positive	positive
Session: all previous vs. re-test	n.s.	-	*negative*	negative
Place of articulation (POA)	positive	-	*positive*	n.s.
POA*session (pre vs. post)	n.s.	-	n.s.	n.s.
POA*session (pre/post-test vs. prod. training)	n.s.	-	n.s.	positive
POA*session (all previous vs. re-test)	n.s.	-	n.s.	n.s.

Table 5: Summary of effects for negative VOT models. Significant effects ($p < 0.05$, as assessed with likelihood ratio tests) are reported in plain text, marginal effects ($p < 0.10$) are italicized. Only the direction of effects is noted; full model estimates are reported in Appendix D. There was an insufficient number of non-zero data points to construct a model for aspirated tokens.

	Pre-test	**Post-test**	**Production training**	**Re-test**
Classification accuracy	51.0%	49.9%	59.1%	52.8%
Permutation tests > classifier	36.9%	51.1%	6.1%	40.8%

Table 6: Classification results and results of permutation tests by session, formant data. Sessions where no more than 5% of the permutation tests meet or exceed the classifier accuracy are considered to be significantly different from chance; 10% represents a marginal threshold.

	Pre-test	Post-test	Production training	Re-test
Classification accuracy	53.9%	55.1%	65.0%	53.4%
Permutation tests > classifier	3.9%	2.2%	0.8%	12.5%

Table 7: Classification results and results of permutation tests by session, formant data. Sessions where no more than 5% of the permutation tests meet or exceed the classifier accuracy are considered to be significantly different from chance; 10% represents a marginal threshold.

Appendix C Figures

Figure 1: Examples of the four voicing categories with dental place of articulation. Intervals labeled with P show positive VOT; intervals labeled with N show negative VOT.

Figure 2: Example screenshots from production training. Figure (A) and (B) show training of the dental-retroflex place contrast by introducing subjects to major articulatory landmarks using sagittal sections. Color cues remind learners of dental (green) and retroflex (red) place of articulation. Figure (C) introduces the concept of aspiration, and a picture to associate with the concept. Figure (D) compares the aspirated "t" to the unaspirated "d" (English orthography). Figure (E) asks subjects to practice combining place and voicing with visual cues. Figure (F) demonstrates a repetition trial for a syllable with the target consonant /ʈʰ/, with visual cues.

Figure 3: Measures of positive VOT. (A) Positive VOT (log-transformed) by voicing category at pre-test, showing long durations for breathy and aspirated stops, and relatively short durations for voiced and voiceless stops. (B) Ratio of later-session durations to pre-test durations of positive VOT, by voicing category, showing changes after training. Errors bars represent standard errors.

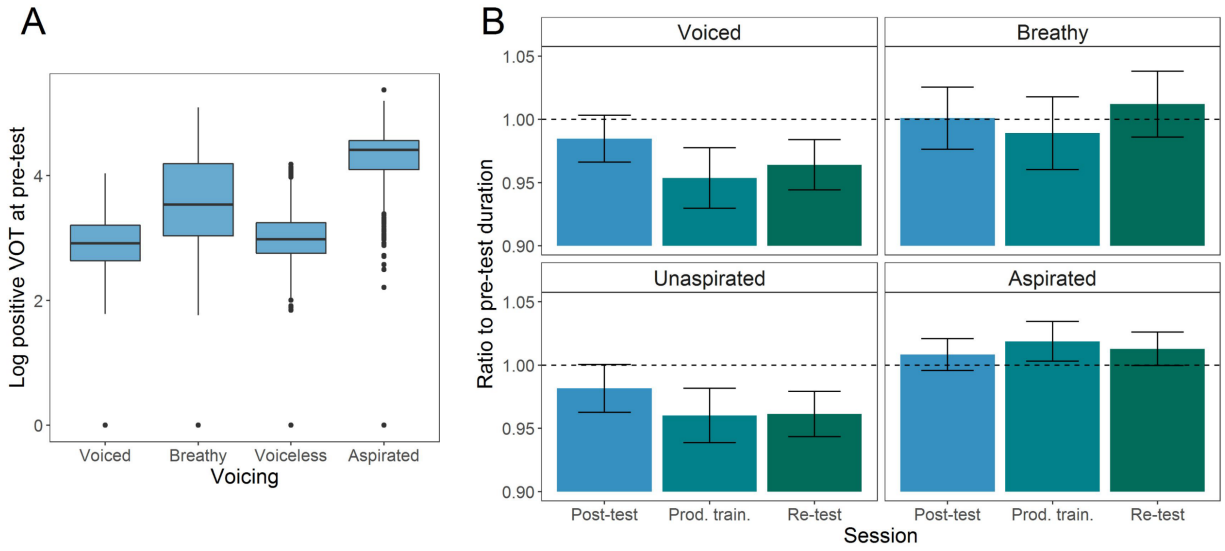


Figure 4: Comparisons of mean negative VOT to pre-test, by follow-up session. Each point represents the mean negative VOT productions from a single speaker of (A) voiced and (B) breathy tokens. Points falling above the diagonal line indicate an increase in negative VOT from pre-test to follow-up. Points falling below the diagonal indicate speakers who reduced their mean VOT duration from pre-test to follow-up. Points at the origin indicate speakers who produced no negative VOT in either session.

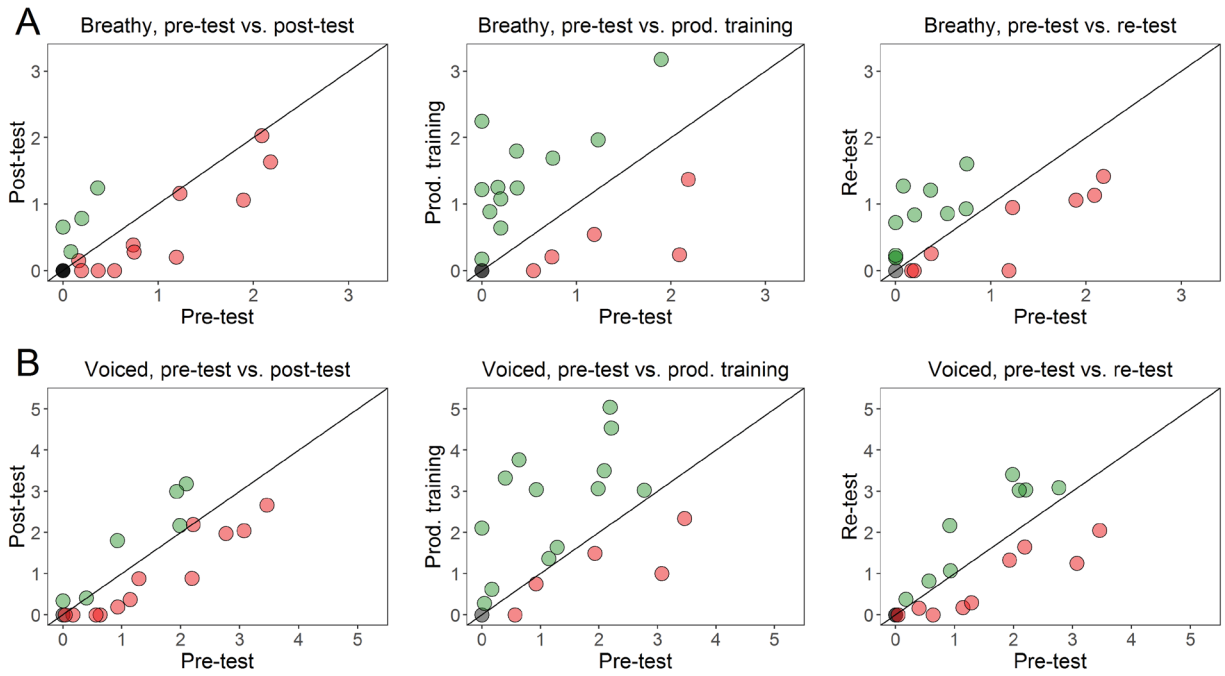
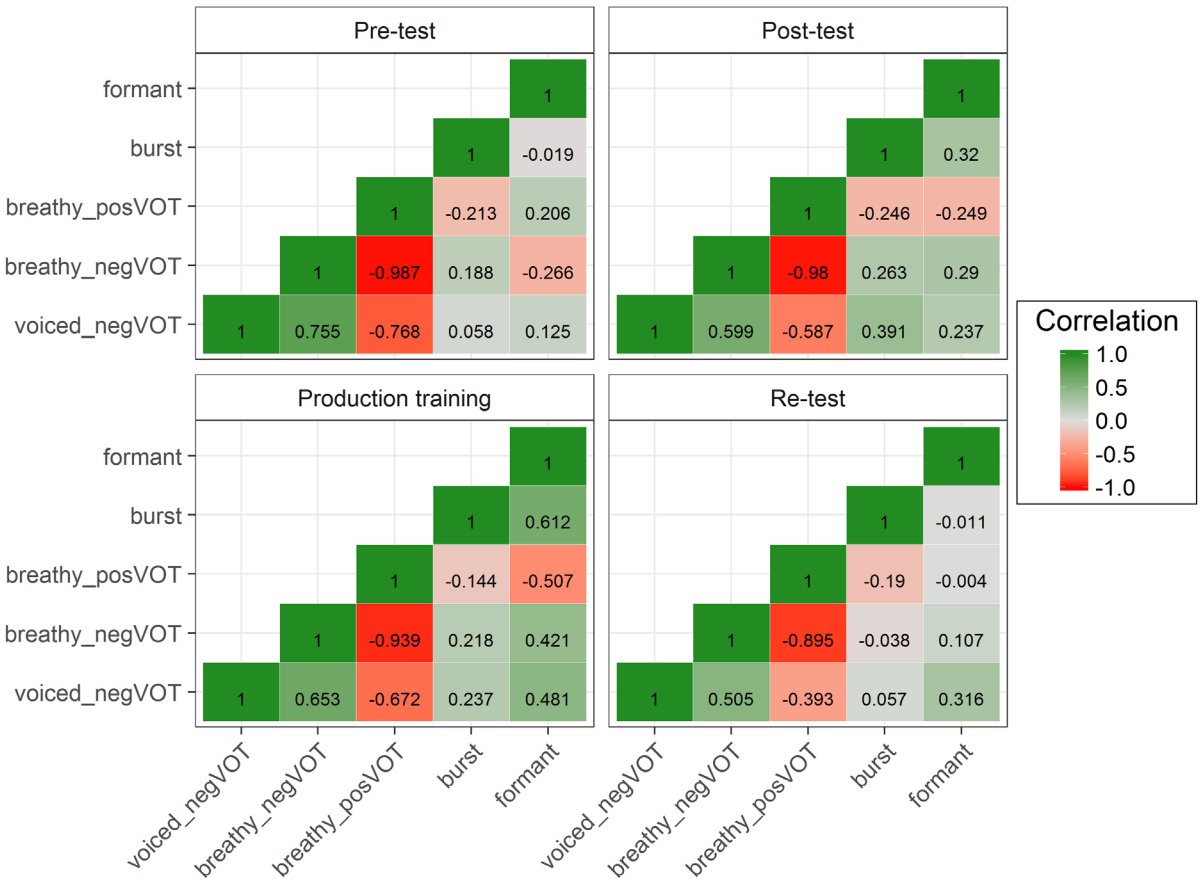


Figure 5: Accuracy analyses by participant, for the (A) formant classification and (B) burst spectrum classification. Each bar shows the mean and standard error of classification accuracy for that session, as assessed by LDA models run on each individual participant's data. Chance (50%) is indicated by the horizontal dashed line. Individual classification accuracy values are plotted in the light grey lines superimposed on each bar.

Figure 6: Correlation matrix of features by session. Positive correlations indicate consistency in production of both features; negative correlations indicate that participants who successfully produced one feature were unsuccessful at the other.



Appendix D VOT model structures and tables

This section contains details on the full model structure and output for the VOT models reported in sections 3.1.2 and 3.1.3. All models had the following fixed effects: mean L2 experience (centered, continuous), mean L3 experience (centered, continuous), session: pre-test vs. post-test (reverse Helmert coded), session: pre/post-test vs. production training (reverse Helmert coded), session: all previous sessions vs. re-test (reverse Helmert coded), place of articulation (dental= -0.5, retroflex = 0.5), and the interaction of place of articulation and all session predictors.

D.1 Positive VOT random effects structure

Unaspirated model, random effects structure: Decorrelated random subject slopes for all session predictors, place of articulation, and the interaction of place of articulation and session (pre/post test vs. production training). Decorrelated random item slopes for place of articulation, session (pre/post-test vs. production training), and the interaction of these two predictors.

Aspirated model, random effects structure: Subject and item intercepts only.

Voiced model, random effects structure: Subject and item intercepts only.

Breathy model, random effects structure: Correlated random subject slopes for all session predictors, place of articulation, and the interaction of place and session (all previous sessions vs. re-test). Correlated random item slopes for L2 experience, L3 experience, and place of articulation.

	Unaspirated	Aspirated	Voiced	Breathy
Constant	2.921 (0.051) t = 57.427***	4.317 (0.065) t = 65.960***	2.838 (0.052) t = 54.062***	3.556 (0.095) t = 37.549***
Mean L2 experience	−0.036 (0.047) t = −0.758	−0.025 (0.062) t = −0.402	−0.023 (0.048) t = −0.473	−0.120 (0.057) t = −2.127
Mean L3 experience	0.044 (0.039) t = 1.118	0.078 (0.051) t = 1.524	−0.008 (0.040) t = −0.199	0.096 (0.046) t = 2.087
Session (pre-test vs. post-test)	−0.107 (0.026) t = −4.074***	0.009 (0.021) t = 0.426	−0.091 (0.025) t = −3.573***	−0.018 (0.051) t = −0.348
Session (pre/post-test vs. prod. training)	−0.131 (0.065) t = −2.028*	0.079 (0.018) t = 4.306	−0.102 (0.022) t = −4.582***	−0.003 (0.082) t = −0.032
Session (all previous vs. re-test)	−0.099 (0.032) t = −3.121***	−0.018 (0.017) t = −1.028	−0.063 (0.021) t = −2.983***	0.052 (0.062) t = 0.810
Place of articulation (PoA)	−0.152 (0.043) t = −3.512***	0.035 (0.032) t = 1.077	−0.088 (0.041) t = −2.168**	−0.050 (0.140) t = −0.357
PoA*session (pre-test vs. post-test)	0.011 (0.045) t = 0.255	−0.057 (0.042) t = −1.358	0.001 (0.051) t = 0.028	−0.039 (0.076) t = −0.521
PoA*session (pre/post-test vs. prod. training)	0.054 (0.057) t = 0.961	−0.041 (0.037) t = −1.117	−0.027 (0.045) t = −0.614	−0.037 (0.098) t = −0.374
PoA*session (all previous vs. re-test)	0.055 (0.037) t = 1.483†	−0.027 (0.035) t = −0.772	0.012 (0.042) t = 0.294	0.035 (0.063) t = 0.556

Total observations	1646	1704	1712	1673

Table 8: Estimates (s.d.) and t-statistics for fixed effects of positive VOT models. Significance was assessed using χ^2 comparisons of nested models with each predictor held out, with * indicating $p < 0.1$, ** indicating $p < 0.05$, and *** indicating $p < 0.01$. The symbol † indicates that the nested model with this predictor held out failed to converge. In these cases, a rough criterion based on the t-statistic was used: effects with $t > 2$ were inferred to be reliable.

D.2 Negative VOT random effects structure

Unaspirated model, random effects structure: Decorrelated random subject slopes for all session predictors and place of articulation. Decorrelated random item slopes for L2 experience, session (all previous sessions vs. re-test), place, and the session by place

interaction.

Voiced model, random effects structure: Decorrelated random subject slopes for all session predictors and place of articulation. Decorrelated item slopes for session (pre/posttest vs. production training) and session (all previous sessions vs. re-test), place of articulation, L2 experience, and L3 experience.

Breathy model, random effects structure: Correlated random subject slopes for all session predictors and place of articulation. Correlated item slopes for place of articulation, L2 experience, and L3 experience.

	Unaspirated	**Voiced**	**Breathy**
Intercept	0.414 (0.133) t = 3.112***	1.474 (0.237) t = 6.231***	0.522 (0.131) t = 3.969***
Mean L2 experience	−0.131 (0.127) t = −1.029†	−0.078 (0.223) t = −0.351	0.069 (0.097) t = 0.708
Mean L3 experience	−0.002 (0.104) t = −0.020	0.139 (0.183) t = 0.760	0.034 (0.079) t = 0.429†
Session (pre-test vs. post-test)	−0.214 (0.197) t = −1.090	−0.193 (0.151) t = −1.272	−0.108 (0.112) t = −0.966†
Session (pre/post-test vs. prod. training)	−0.104 (0.101) t = −1.031	0.924 (0.353) t = 2.619**†	0.359 (0.205) t = 1.751*†
Session (all previous vs. re-test)	0.018 (0.142) t = 0.126†	−0.331 (0.167) t = −1.979*†	−0.241 (0.112) t = −2.152**
Place of articulation (PoA)	0.375 (0.151) t = 2.477**	0.29 (0.169) t = 1.715*	0.097 (0.149) t = 0.651†
PoA*session (pre-test vs. post-test)	0.114 (0.129) t = 0.881	−0.045 (0.230) t = −0.196	0.368 (0.159) t = 2.317**

PoA*session (pre/post-test vs. prod. training)	-0.072 (0.113) t = −0.639	0.330 (0.221) t = 1.493	0.092 (0.142) t = 0.646†
PoA*session (all previous vs. re-test)	-0.010 (0.108) t = −0.096	-0.207 (0.198) t = −1.046	-0.222 (0.135) t = −1.638†
Total observations	1,603	1,726	1,625

Table 9: Estimates (s.d.) and t-statistics for fixed effects of negative VOT models (note: no aspirated model was run due to insufficient variance in the dependent variable). Significance was assessed using χ^2 comparisons of nested models with each predictor held out, with * indicating $p < 0.1$, ** indicating $p < 0.05$, and *** indicating $p < 0.01$. The symbol † indicates that the nested model with this predictor held out failed to converge. In these cases, a rough criterion based on the t-statistic was used: effects with $t > 2$ were inferred to be reliable, and > 1.7 to be marginal.