Does voicing affect patterns of transfer in non-native cluster learning?

Hung-Shao Cheng and Adam Buchwald
Department of Communicative Sciences and Disorders
New York University

New York, NY

Conflict of Interest statement:
There are no conflict of interests

Funding statement:
This work was supported by a grant awarded from National Institute on Deafness and other Communication Disorders (NIDCD) to Adam Buchwald (K01DC014298).

Corresponding author:
Hung-Shao Cheng
Department of Communicative Sciences and Disorders
New York University
665 Broadway, Suite 900
New York, NY 10012
Email: hscheng@nyu.edu

ABSTRACT

Purpose. Previous studies have demonstrated that speakers can learn novel speech sequences, although the content and specificity of the learned speech-motor representations remain incompletely understood. We investigated these representations by examining transfer of learning in the context of non-native consonant clusters. Specifically, we investigated whether American English speakers who learn to produce either voiced or voiceless stop-stop clusters (e.g. /gd/ or /kt/) exhibit transfer to the other voicing pattern.

Method. Each participant (n = 34) was trained on disyllabic nonwords beginning with either voiced (/gd/, /db/, /gb/) or voiceless (/kt/, /kp/, /tp/) onset consonant clusters (e.g., /gdimu/, /ktaksnæm/) in a practice-based speech motor learning paradigm. All participants were tested on both voiced and voiceless clusters at baseline (prior to practice) and in two retention sessions (20 minutes and 2 days after practice). We compared changes in cluster accuracy and burst-to-burst duration between baseline and each retention session to evaluate learning (performance on the trained clusters) and transfer (performance on the untrained clusters).

Results. Participants in both training conditions improved with respect to cluster accuracy and burst-to-burst duration for the clusters they practiced on. A bidirectional transfer pattern was found, such that participants also improved the cluster accuracy and burst-to-burst duration for the clusters with the other untrained voicing pattern. Post hoc analyses also revealed that improvement the production of untrained stop-fricative clusters that originally were added as filler items.

Conclusions. Our findings suggest the learned speech motor representations may encode the information about the coordination of oral articulators for stop-stop clusters independently from information about the coordination of oral and laryngeal articulators.

KEYWORDS: Speech motor learning, non-native clusters, transfer of learning

Introduction

Speech production is a complex motor behavior that involves precise spatiotemporal control and coordination of the speech articulators to produce linguistically meaningful sequences. While executing the speech motor sequences in one's native language may be effortless, it can be more difficult to learn to produce novel sequences. Understanding how this learning occurs can provide insight into understanding speech motor learning more generally. Previous studies have investigated speech motor learning in neurotypical adult speakers with various non-native speech targets including singleton consonants (Katz & Mehta, 2015; Levitt & Katz, 2007), consonant clusters (Buchwald et al., 2019; Segawa et al., 2019; Segawa et al., 2015; Steinberg Lowe & Buchwald, 2017), and vowels (Carey et al., 2017; Kartushina et al., 2015; Kartushina et al., 2016; Kartushina & Martin, 2019; Li et al., 2019). While these studies have consistently reported improvement in the production of the trained non-native speech targets, the content and specificity of these learned speech motor representations remain incompletely understood.

Given that the specificity of speech sound representations cannot be understood by examining improvement on trained targets alone, the extent to which the learning transfers to other (untrained but related) speech motor targets has been used to understand what is encoded in the learned speech motor representation (Maas et al., 2008). When transfer occurs, we may assume that the representations governing the production of the two items share enough content to allow the learning to affect both items. Understanding the patterns of transfer can then be used to enhance the effectiveness and efficiency of speech motor learning-based treatment by optimizing the selection of training targets to have the broadest improvement. The aim of the present study is to evaluate whether voiced and voiceless non-native consonant clusters share the same learned representation. We trained neurotypical adult speakers of American English on either voiced or voiceless stop-stop clusters (e.g., voiced: /gd/ as in /gdi.vu/; voiceless: /kt/ as in /ktɑ.mi/), and examined their production of the trained items, generalization to untrained items containing the trained cluster, and transfer to the other untrained voicing category. In the following section, we describe how using a transfer paradigm in this context may allow us to better understand speech motor representations.

Transfer in speech motor learning

In this paper, we use the term *generalization* to refer to the ability to produce the same learned speech sound sequence (for example, non-native consonant cluster) in a novel word, and we use the term *transfer* to refer to the ability to produce an untrained speech sound sequence. In previous studies examining transfer of learning, varying approaches have been used to examine the extent to which learning on one item transfers to performance on another item. In one set of studies that focuses on speech sensorimotor adaptation (e.g., Houde & Jordan, 1998), speakers are asked to produce a target speech sound and are provided with real-time sensory feedback (e.g., auditory or somatosensory) of their own production. A perturbation is introduced in either the auditory or somatosensory feedback, and learning is operationalized as the extent to which speakers adapt to the perturbation. In this paradigm, transfer is assessed based on the amount of adaptation found on untrained speech sounds when the perturbation is removed. In many studies, transfer was found to be dependent on acoustic or articulatory similarity between trained and untrained vowels (Cai et al., 2010; Caudrelier et al., 2018; Houde & Jordan, 1998; Rochet-Capellan et al., 2012; but see Tremblay et al., 2008), suggesting that the specific acoustic and articulatory information of the trained vowel is encoded in the learned representation after sensorimotor adaptation.

In another set of studies targeting speech motor treatment in individuals with apraxia of speech (Austermann Hula et al., 2008; Ballard, 2001; Ballard et al., 2007; Knock et al., 2000; Wambaugh et al., 1998), speakers receive treatment targeting specific speech sounds and then researchers examine whether they improve at producing those sounds, and whether the improvement transfers to untreated sounds. The preliminary findings from this domain indicate that training sounds involving one manner of articulation (e.g., stops) can transfer to other sounds in that class but not to sounds involving a different manner of articulation (e.g., fricatives) (Ballard et al., 2007; Knock et al., 2000; Wambaugh et al., 1998). The results have been interpreted as indicating that transfer does not occur across different manners of articulation, and therefore that speech motor representations of consonants encode manner. However, most studies have primarily examined transfer across different manners of articulation; the degree to which transfer can occur between different voicing categories with the same manner of articulation remains incompletely understood.

Taken together, the above studies suggest that there are clear constraints on how transfer occurs within speech motor learning, and these are taken to reflect the nature of the learned representations. To the best of our knowledge, whether transfer can occur between voicing categories has not been explicitly examined. Therefore, the current study aims to address this question in the context of non-native consonant cluster learning.

Non-native consonant cluster production and learning

The successful production of onset consonant clusters is characterized by a precise gestural coordination pattern among the articulators involved (Browman & Goldstein, 1988, 1995; Byrd, 1996) although the exact coordination pattern differs across consonant types and languages (Chitoran et al., 2002; Marin & Pouplier, 2010; Pastätter & Pouplier, 2017; Pouplier et al., 2017). In terms of voicing control in consonant clusters, the gesture of the oral articulators needs to be tightly coordinated with the gesture of laryngeal articulator in order to manifest the correct voicing pattern (Bombien & Hoole, 2013; Hoole & Bombien, 2014; Hoole & Bombien, 2017; Löfqvist, 1980; Löfqvist & Yoshioka, 1980, 1984). While onset consonant clusters are permitted in English (Marin & Pouplier, 2010), stop-stop clusters are phonotactically illegal in syllable initial position. In their study on non-native onset cluster production with American English speakers, Davidson (2010) reported that the most frequent error type in producing stop-stop onset clusters is vowel epenthesis in between the two consonants. This error is further thought to arise due to the mis-timing of the gestural coordination of individual consonant productions. Thus, the gestural timing between the articulators may represent the potential phonetic target to learn for American English speakers.

Previous studies of non-native cluster learning have suggested that learning occurs at the level of non-native clusters instead of at the item level (Buchwald et al., 2019; Segawa et al., 2019). For example, Buchwald et al. (2019) investigated learning on a wide range of non-native onset clusters (e.g., /zb/, /vm/) embedded in disyllabic nonwords (e.g., /zbu.kip/, /vmæ.ki/) in adult American English speakers without impairment as part of a larger study on neuromodulation. The behavioral results of their study indicated that participants who were trained to produce onset clusters in four nonwords showed increased accuracy of the trained onset clusters in both the trained nonwords and untrained nonwords that contained the trained clusters. This suggests that speakers learn to produce the non-native cluster, not just a specific item (also see Segawa et al., 2019).

While the above studies demonstrated training on an onset consonant cluster in some nonwords can generalize to other nonwords with the same cluster, the extent to which learning to produce a novel onset consonant cluster can transfer to other untrained non-native consonant clusters remains largely unexplored. We explore this context with respect to clusters that involve a non-native onset consonant sequence in English (stop-stop clusters) and differ in their voicing status. Thus, the oral-to-oral articulatory coordination for the two cluster types are similar, but they involve a different oral-to-laryngeal coordination. The next section outlines our approach and the specific research questions that motivated our experimental work.

The current study: Transfer of learning across voicing categories in stop-stop onset clusters

The current study aimed to investigate whether transfer of learning can occur across voicing categories. In particular, we trained speakers to produce items beginning with either voiced stop-stop clusters (/gd/, /gb/, /db/) or voiceless stop-stop clusters (/kt/, /kp/, /tp/). We used a practice-based speech motor learning paradigm that included a pre-practice component, in which participants received general instructions on how to produce consonant clusters so they knew what the target was during practice, and a practice component based on parameters reported to enhance motor learning (Maas et al., 2008). Participants were tested on both sets of clusters at baseline and again at two retention points. Within each trained cluster, we trained on some nonwords and tested on others to explicitly replicate the finding that training on non-native clusters in some nonwords generalizes to the production of the same clusters in untrained nonwords (research question 1 below). Our primary focus was to test whether learning one class of stop-stop cluster (voiced or voiceless) can transfer to the other class (research question 2 below). If learned representations encode the coordination pattern between oral-to-oral articulators of the clusters regardless of voicing, we would expect a bidirectional transfer pattern, with each training group improving at both types of clusters. We would take this finding to indicate that the representation of the speech-motor plan for producing stop-stop clusters encodes information about oral-to-oral articulator coordination separately from information about the laryngeal articulators and the oral-laryngeal coordination; thus, what is learned about the oral articulators can transfer across these categories. Conversely, if the coordination pattern between oral-to-oral articulators is encoded together with the information regarding the laryngeal articulators, we would not expect transfer between voicing categories.

Another factor that may affect transfer of learning is the complexity of speech motor representations, with the idea that learning more complex patterns may transfer to the less complex ones, but not vice versa (Maas et al., 2008). While complexity has been investigated often in studies of speech motor control (Riecker et al., 2008; Sadagopan & Smith, 2008), relatively little work explicitly addressing how complexity interacts with transfer of learning. Within this narrower domain, the effect of complexity has primarily been investigated in individuals with acquired apraxia of speech, and has yielded equivocal findings (Maas et al., 2002; Schneider & Frens, 2005), although we note that the idea of training more complex targets to promote transfer of learning has been influential in other domains involving speech and language rehabilitation (e.g., Thompson et al., 2003). Thus, we considered the possibility that complexity would affect transfer. We considered voiced clusters to be more complex than their voiceless counterparts for both phonological and phonetic reasons. Phonologically, voiceless clusters are considered less *marked* based on their cross-linguistic distribution (Morelli, 1999); the existence of voiced clusters in a language predicts the existence of voiceless counterparts, whereas the reverse is not true. Phonetically, aerodynamic studies have suggested that it is difficult to maintain phonation during closure as required in the production of voiced stop-stop clusters, whereas the production of voiceless stop-stop clusters does not require the phonation during closure (Kawasaki-Fukumori & Ohala, 1997; Ohala, 1983, 1997). In addition, previous studies on non-native cluster production have reported lower accuracy for voiced stop-stop clusters than voiceless stop-stop clusters (Davidson, 2006, 2010; Wilson et al., 2014). Thus, if complexity plays a role in transfer of learning, we would expect to see an asymmetrical transfer pattern, with learning on the more complex voiced clusters transferring to untrained voiceless clusters more than training on voiceless clusters would transfer to the voiced clusters (research question 3 below).

In summary, the present study was designed to address the following questions:

1) Does training on voiced or voiceless stop-stop clusters in some nonwords generalize to untrained nonwords that contain the trained clusters?

2) Does training on voiced or voiceless stop-stop clusters transfer to nonwords that contain clusters with the untrained voicing specification?

3) Is there a difference in the magnitude of the transfer effect when training on voiced stop-stop clusters vs. voiceless stop-stop clusters?

We note here that we included a smaller number of additional non-native clusters as filler items (stop-fricative onset clusters) that were not initially intended to be part of these research questions. Based on the findings of the primary questions, we also analyzed changes in production on these clusters as well, as described in the methods and results.

Methods

Participants

Thirty-four neurotypical adult participants (11M, 23F; mean age = 23.8 yrs) completed the study. All participants were native speakers of American English. Participants were excluded if they reported a history of speech, hearing, or neurological disorder; if they were familiar with languages that contained stop-stop clusters that are used in this study, such as Russian, Polish, Czech, Greek, Arabic, and Hebrew; or if they had any prior training in phonetics or speech science. All participants reported normal or corrected-to-normal vision and all passed an oral-motor examination and a pure-tone hearing screening (25 dB at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz). Informed consent was obtained according to the NYU institutional review board. Participants received compensation ($25) at the end of the second day of the experiment. An additional eleven adult participants were initially consented but did not complete the entire experiment: seven had failed to disclose in email screening that they met exclusion criteria (two for language background requirement and five for history of speech disorders), technical issues with computer software ruled out three participants, and one failed to return for the second retention session.

Speech stimuli

The target stimuli were disyllabic nonwords beginning with either voiced stop-stop or voiceless stop-stop onset clusters (e.g., /gdum.prid/, /ktɑk.snæm/; See Appendix A for full list of stimulus words). Six target clusters were used: three voiced stop-stop clusters (/gd/, /gb/, /db/), and three voiceless stop-stop clusters (/kt/, /kp/, /tp/). Eight distinct nonwords were recorded for each of these six clusters. The syllable shape for each target nonword varied with respect to its CV (consonant-vowel) structure, and the nucleus of the first (stressed) syllable was either /i/, /ɑ/, or /u/. We also included 27 filler nonword stimuli during baseline and retention sessions to increase the variability of the task, including items with singleton onsets, phonotactically legal consonants clusters (e.g., /sn/, /sm/), and phonotactically illegal stop-fricative onset clusters (e.g., /gz/, /kf/) (See Appendix A2). The phonotactically illegal stop-fricative stimuli were designed to match the stop-stop items with respect to syllable structure and place of articulation of the consonants.

All speech stimuli were recorded by a phonetically-trained Polish-American English simultaneous bilingual speaker using a Shure SM-10 head-mounted microphone attached to a Marantz PMD660 digital recorder. All sound files were spliced to leave 60 ms of silence at the onset of each item. The files were then down-sampled to 22050 Hz and normalized to the mean amplitude of all sound files using Praat (Boersma & Weenink, 2019). Orthographic versions of the nonwords were created according to American English orthography and were verified by native speakers of American English to ensure they elicited the correct grapheme-to-phoneme correspondences.

Procedure

All components of the experiment took place in a sound-attenuated testing room. Participants were seated in front of a computer and their productions were recorded using a Shure BETA 58A microphone in a desktop microphone stand connected to the Marantz PMD660 digital recorder. The experiment was implemented in PsychoPy (Pierece, 2007). The overall structure of the procedure is presented in Figure 1. Participants were randomly assigned to either the voiced or voiceless cluster training group prior to beginning the study. We first describe the components of the speech motor learning paradigm, and then the additional tasks that were performed.

Baseline. The baseline session began after participants were consented. During the baseline, participants repeated the items described above that were presented both auditorily and orthographically. Each trial began with a fixation cross for 250 ms, followed by a blank screen for 150 ms. The orthography was then presented and remained on the screen for 2050 ms. The auditory model began 50 ms after the onset of the orthography. The screen then remained blank until the onset of the fixation cross for the next trial. Participants were instructed to respond as soon as they were ready after the auditory model was finished playing. The participants produced all eight nonwords per cluster (48 unique nonwords) twice each. In addition, participants produced the 27 filler words twice each. The stimuli were randomized and presented in two blocks. The baseline session lasted approximately 15 minutes with no feedback provided.

Pre-practice. The pre-practice began immediately after the baseline session. The goal of the pre-practice was to ensure that participants understood the targets they were supposed to practice. First, the idea of how clusters contrast with singletons was introduced using the example word pair “bleed” and “believe.” Then participants were presented with two items with non-native clusters that were not part of the present study (/ftɑ.næd/, /fmi.du/) and asked to produce them twice each. After each repetition, we reiterated that the onset consonant clusters should be produced with the consonant sounds ‘together,’ without putting a vowel in between the two consonant sounds. The pre-practice session lasted approximately 2 minutes.

Practice. During the practice session, participants were instructed to use their pre-practice training to repeat nonwords following simultaneous auditory and orthographic models, with the same timing as the baseline session. Participants produced exclusively voiced or voiceless stop-stop sequences depending on their random group assignment. Each participant repeated 4 nonwords per target stop-stop cluster ten times each (120 total). The target nonwords were counterbalanced across participants within each of the practice conditions, such that half of the participants were trained on one half of the nonwords and the other half on the second half of the nonwords. In addition, participants produced a total of 60 additional phonotactically-legal nonwords with singleton onsets (i.e., /r/, /l/, /w/) and legal English onset clusters (i.e., /bl/, /sm/,

/fr/). The practice session was structured to be consistent with several principles of motor learning that enhance learning (Maas et al., 2008). In particular, we included a large number of trials and the stop-stop clusters were presented in variable phonetic contexts. In addition, the stimuli were pseudo-randomized to ensure that no same target cluster was presented in succession and that no nonword occurred twice within three trials. Because of the difficulty of perceiving these clusters for speakers of languages that do not contain the clusters (Davidson, 2006, 2007), we did not provide any feedback regarding the production accuracy to the participants during the practice session. The practice session lasted approximately 20 minutes.

Retention sessions. The first retention (R1) and the second retention (R2) were structured identically to the baseline session. The first retention took place 20 minutes after the practice session, with a series of tasks performed during this time (see below). Participants returned to the lab two days after the first session for R2. As in the baseline, no feedback was provided regarding production accuracy. Each retention session lasted approximately 15 minutes.

Additional tasks. Prior to the baseline, participants were given verbal (i.e., forward and backward digit span) and visuo-spatial (forward and backward block span) working memory tasks. These data were not analyzed in the current study. To ensure at least 20 minutes passed between the practice and retention sessions, we designed a small battery of tasks to be given during this time. Participants were given the pure-tone hearing screening test described in the Participants section. In addition, the diadochokinetic (DDK) syllable repetition task as well as an oral-motor examination was performed to ensure participants' oral-motor abilities were within functional limits.

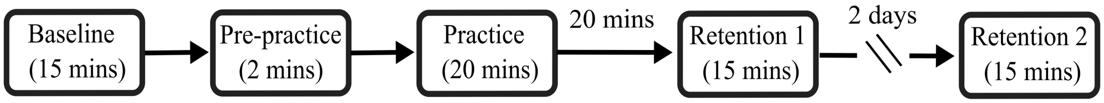

Figure 1 A schematic representing the procedure of the training paradigm

Data analysis

Cluster accuracy

For each participant, the full set of recordings were divided into smaller units and randomized for the purpose of blinding the raters to the experimental session. The recordings were coded by two raters who were blind to the participant's training conditions (i.e., voiced or voiceless) and to the experimental sessions (i.e., baseline, R1, and R2). All recordings were coded using Praat (Boersma & Weenink, 2019). For cluster accuracy, the most common participant error involves vowel epenthesis (e.g., /gbimu/ → [gəbimu]) (Wilson et al., 2014). Given the aforementioned difficulty of accurately perceiving these sequences, all accuracy measures were based on the presence of a vowel in the acoustic record. Following Wilson et al. (2014) and Buchwald et al. (2019), the presence of a vowel was determined based on two criteria: 1) the presence of (at least) two repetitive vocoid cycles in the acoustic waveform; and 2) the presence of higher formant structures (e.g., F2 and F3) in the spectrogram. Figure 2 depicts two productions of the first syllable in [gbimu], produced without (Figure 2A) and with an epenthetic vowel (Figure 2B).

Other error types, such as deletion (e.g, /gbimu/ → [bimu]), substitution (e.g, /gbimu/ → [grimu]), and metathesis (e.g., /gbimu/ → [bgimu]) and voicing (/gbimu/→[kpimu]) were determined based on a combination of perception and the acoustic record. Cluster accuracy was coded as binary, but the items with other errors were excluded from additional analyses described below. Inter-rater reliability was evaluated on 20% of the data coded by two independent raters and the point-to-point inter-rater agreement was 91%.

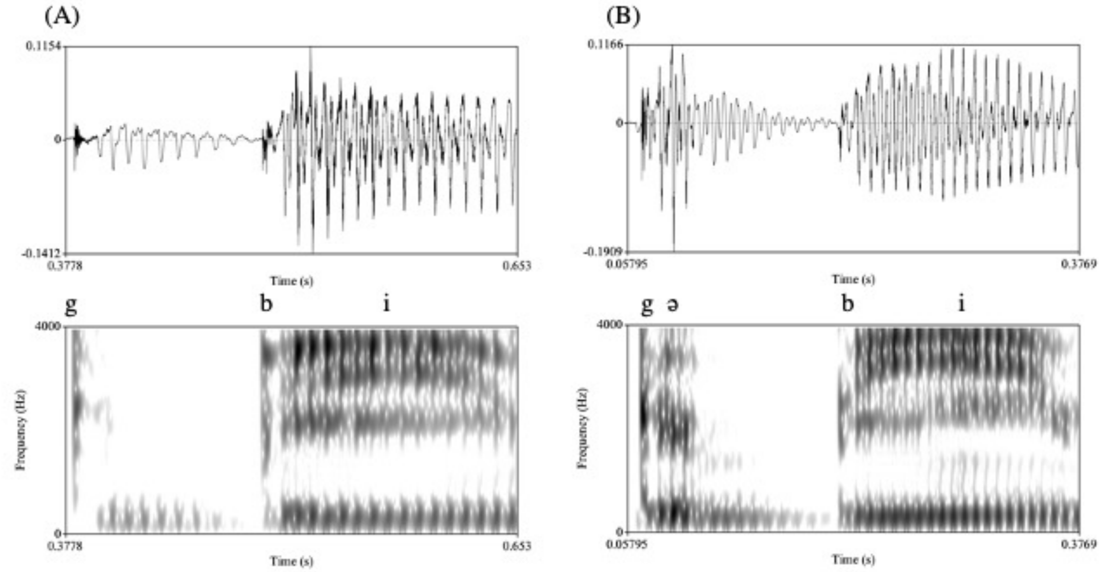

Figure 2. Acoustic waveform and spectrogram of the [gbi] portion in two tokens of [gbimu]. (A) The token was produced without an epenthetic vowel. (B) The token was produced with an epenthetic vowel.

Burst-to-burst duration

Burst-to-burst duration of the stop-stop cluster was measured to examine whether there was a gradual shortening towards a more target-like production based on the training. Only clusters that were either produced correctly or produced with an epenthetic vowel were included in the analysis. We included all tokens where the speaker produced the two consonants at the beginning of the word for two reasons. First, in producing voiceless stop-stop clusters, a speaker may produce the oral articulator patterns associated with an epenthetic vowel, but an absence of phonation would lead this to be unobservable on the acoustic record. In addition, we are using burst-to-burst duration as a continuous measure to evaluate changes in motor acuity, and we want to include the full range of coordination among the oral articulators to determine whether improvement is observed rather than treat this duration as part of a categorical measure. The burst-to-burst duration measured the onset of the acoustic burst of the first stop to the onset of the acoustic burst of the second stop. The onset of the burst was defined as the first zero crossing point after the first trough of the acoustic burst. Since it is common for velar stops to have more than one visible acoustic burst (Repp & Lin, 1989), the last acoustic burst was used. Inter-rater reliability was evaluated on 20% of the data coded by two independent raters, with agreement evaluated based on whether the two measurements were within 10 milliseconds (point-to-point inter-rater agreement: 96%). In addition, because participants produced the nonwords repetitively through the whole experiment, they become more familiar with the nonwords. Thus, a change in burst-to-burst duration could also come from a global increase in the speaking rate. To determine whether the burst-to-burst duration changes came from rate changes, we also measured the duration of the stressed vowel (i.e., the vowel in the first syllable of our disyllabic stimuli) as a proxy for speaking rate, as shown in Figure 3.

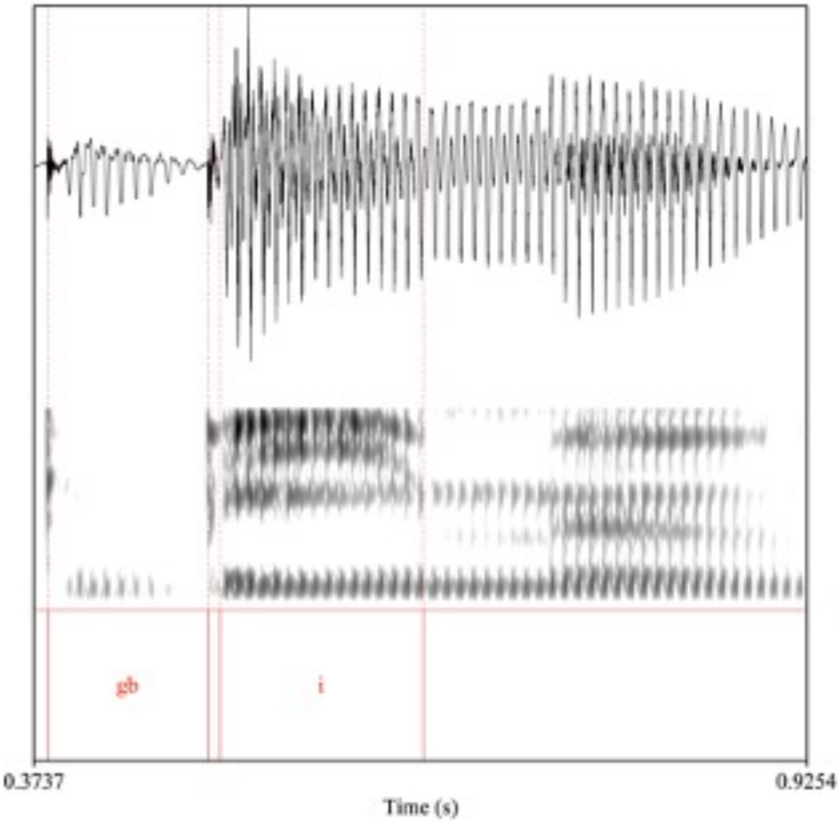

Figure 3. The coding of burst-to-burst duration and vowel duration in Praat. This is the same token as shown in Figure (2A). The onset of the burst was defined as the zero crossing point after the first trough on the waveform.

Statistical analysis

We evaluated speech motor learning by comparing performance for each retention session to the baseline. Separate statistical models were built to analyze cluster accuracy as well as burst-to-burst duration. Within each model, the factor of Training encoded items as Trained (specific tokens used in Practice session), Generalization (untrained items beginning with trained cluster),Transfer (items beginning with untrained clusters). All statistical analyses were conducted in R (R Core Team, 2017). Linear mixed-effects models were implemented by using the *lme4* package (Bates et al., 2015). Data organization and plotting of the results were done by packages *tidyr (*Wickham & Henry, 2019*)*, *dplyr* (Wickham et al., 2019), and *ggplot2* (Wickham, 2016). Cluster accuracy was evaluated using logistic mixed-effects models because the dependent variable is binary; burst-to-burst duration was evaluated using linear mixed-effects models. For each comparison, models began with random intercepts for Participant and Item. Following the statistical approach in Harel and McAllister (2019), we selected the best random effect structure based on AIC (Akaike, 1974) and BIC (Schwarz, 1978) scores. When AIC and BIC scores differed, we selected the model chosen by BIC score for the ease of model interpretation because BIC prefers simpler models than AIC.

For the cluster accuracy analysis, to assess whether there was improvement between the baseline and each of the retention sessions, the logistic mixed-effects model included Condition (Voiced vs. Voiceless), Session (Baseline, R1, R2), and Training (Trained vs. Generalization vs. Transfer), and their interaction as fixed-effect predictors, and the random effects structure preferred by BIC. The model was dummy coded (Davis, 2010), and run with Baseline as the Session reference level so that the baseline session was compared separately to each retention session. To evaluate simple effects of Session on each training group independently, we ran the model with each level of the Training variable set as the reference level for each condition. This approach allowed us to inspect the model for simple effects of improvement. To evaluate the possibility that there were different magnitudes of improvement for each type of item, we examined the interaction of Session and Training. In addition, as one of the research questions (research question 3) pertained to the potential difference in the amount of transfer between Voiced and Voiceless condition, the three-way interaction term of Condition, Session, and Training was included in the model to address this question.

For the burst-to-burst duration analyses, the linear mixed-effects model included Condition (Voiced vs. Voiceless), Session (Baseline, R1, R2), Training (Trained vs. Generalization vs. Transfer), and their interaction as fixed-effect predictors, as well as the vowel duration for each item. Once again, the random effects structure preferred by BIC was included in the linear mixed-effects models. The same statistical approach was used to examine simple and interaction effects in burst-to-burst duration as described above.

Post hoc data analysis: Stop-fricative onset clusters

When participants received training specifically on either voiced or voiceless stop-stop clusters in the practice session, they were given general instructions on how to produce consonant clusters as part of ensuring that they know what targets to practice. Because of this, it remains possible that any generalization and transfer to untrained clusters could arise from this instruction. To properly address this question, we would require a control group that was not trained on stop-stop clusters. In the absence of that group, we further evaluated the performance on the set of items beginning with stop-fricative clusters, which were included as filler items and designed to be similar to the stop-stop targets. As this was a post hoc analysis and had not been part of the design, there were only 36 items for each participant per session (as opposed to 96 items for the stop-stop clusters per session). Cluster accuracy for stop-fricative clusters was coded using the same coding procedure as mentioned above. The recordings were coded by three raters who were blind to participant’s training conditions and to the experimental sessions. To examine whether there is improvement in the fine-grained coordination pattern in stop-fricative clusters, the onset of burst to the offset of the fricative was measured (henceforth, C1-C2 duration). Following Davidson and Roon (2008), the offset of the fricative was defined as the beginning of the formant structure of the following vowel. It is worth mentioning that this duration measure is different from the burst-to-burst duration for the stop-stop clusters, where it examined the interval between the onsets of the two stop bursts. The interval between the onset of the burst to the offset of the fricative was selected because of the difficulty locating the onset of the fricative in the acoustic record.

As with the stop-stop clusters, only tokens that were either produced correctly or produced with an epenthetic vowel were analyzed. The duration of the following vowel was measured as a proxy for speaking rate as well. The same statistical approach as described for the stop-stop clusters was used to model the cluster accuracy and C1-C2 duration for stop-fricative clusters. For cluster accuracy, the mixed-effect logistic model included Condition (Voiced vs. Voiceless), Session (Baseline, R1, R2), Voicing (Voiced vs. Voiceless), and their interaction terms as fixed-effect predictors, as well as the random effect structure preferred by BIC. For C1-C2 duration, the linear mixed-effect model included Condition (Voiced vs. Voiceless), Session (Baseline, R1, R2), Voicing (Voiced vs. Voiceless), and their interaction terms as fixed-effect predictors. In addition, vowel duration for each item was added as fixed-effect predictor. The random effect structure preferred by BIC was included in the model. The data and scripts can be found in our OSF repository (https://osf.io/27ntw/).

Results

Cluster accuracy

Figures 4 and 5 present the cluster accuracy data from the Voiced training (Figure 4) and Voiceless training (Figure 5) conditions respectively. As can be clearly seen in these figures, participants were more accurate at producing voiceless clusters than voiced clusters, regardless of training group. This reflects the underlying difference between these clusters with respect to motor implementation, as the voiced clusters require coordination between the oral and laryngeal articulators as well as similar coordination within the oral vocal tract. We consider the statistical outcomes relevant to the primary research questions of this paper here and revisit this observation in the Discussion.

AIC and BIC preferred the model that included random intercepts for participant and item. The model revealed that, for the Voiced training condition, the accuracy for Trained voiced clusters significantly improved from Baseline to both R1 ($\beta = 0.87$, SE = 0.19, $p < .0001$) and R2 ($\beta = 0.51$, SE = 0.19, $p = .008$). This same pattern of improvement was seen for the Generalization items, which improved from Baselined to R1 ($\beta = 0.77$, SE = 0.19, $p < .0001$) and R2 ($\beta = 0.41$, SE = 0.19, $p = .03$), as well as for the Transfer (voiceless cluster) items (R1 vs. Baseline: $\beta = 0.99$, SE = 0.12, $p < .0001$; R2 vs. Baseline: $\beta = 0.49$, SE = 0.12, $p < .0001$) (Figure 4). None of the interactions between Session and Training were significant. The results revealed that participants who practiced voiced clusters improved at trained items, generalized that learning to untrained items with those clusters, and transferred the learning to voiceless clusters.

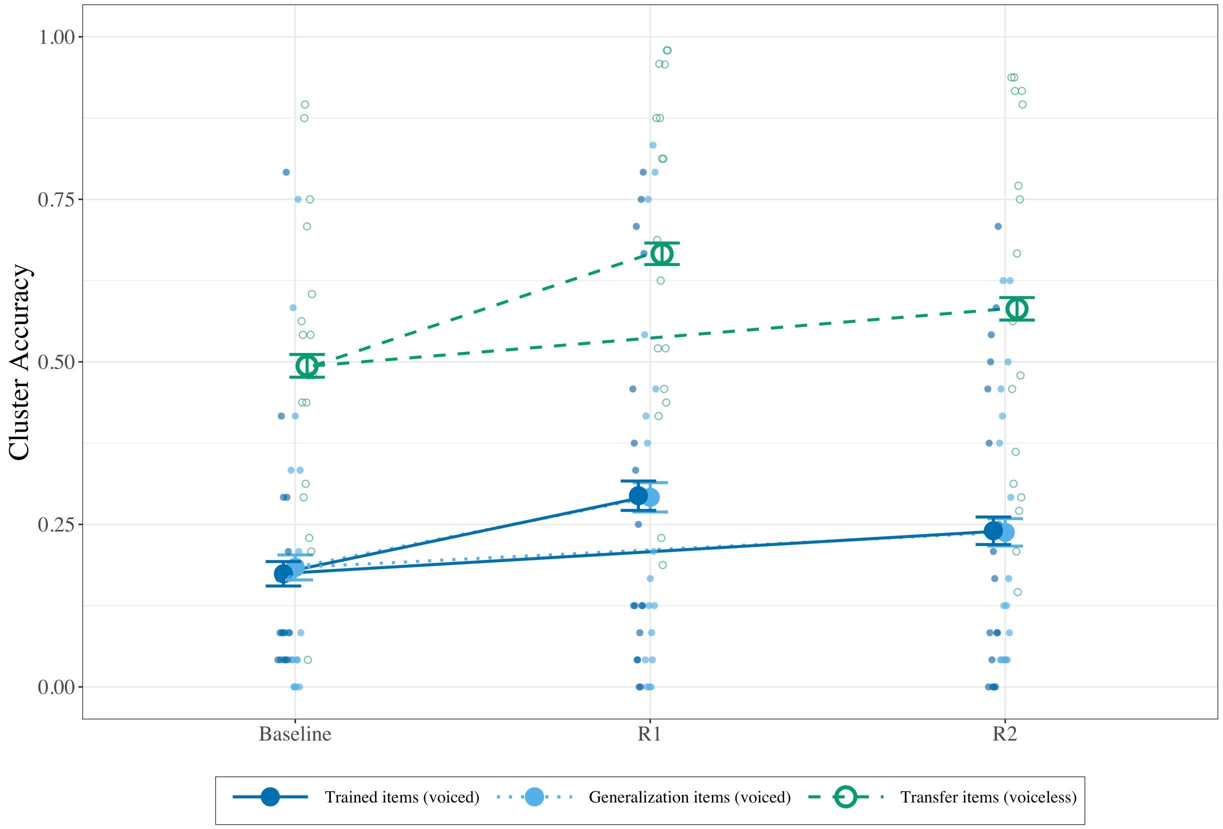

Figure 4. Change in cluster accuracy for the voiced training condition. The figure depicts overall cluster accuracy for each stimulus group from baseline to R1 and R2. The mean group accuracy was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

For the Voiceless training condition, the model revealed that there was significant improvement from the baseline to each retention session for the Trained items (R1: $\beta = 1.15$, SE = 0.19, $p < .0001$; R2: $\beta = 0.74$, SE = 0.18, $p < .0001$), Generalization items (R1: $\beta = 1.44$, SE = 0.19, $p < .0001$; R2: $\beta = 1.03$, SE = 0.18, $p < .0001$), and Transfer items (R1: $\beta = 1.51$, SE = 0.14, $p < .0001$; R2: $\beta = 0.86$, SE = 0.14, $p < .0001$) (Figure 5). Once again, there were no significant interactions between Session and Training. Moreover, there was no significant three-way interaction between Condition, Session and Training. Taken together, the findings regarding cluster accuracy revealed that participants improved in their accuracy on the trained items, they generalized their learning to untrained nonwords with those clusters, and this learning transferred to the other cluster. The lack of any significant interactions in the model demonstrates that the amount of improvement on trained items was not statistically different from the improvement on either generalization or transfer items. Additionally, the amount of generalization and transfer did not differ between the Voiced and the Voiceless training conditions.

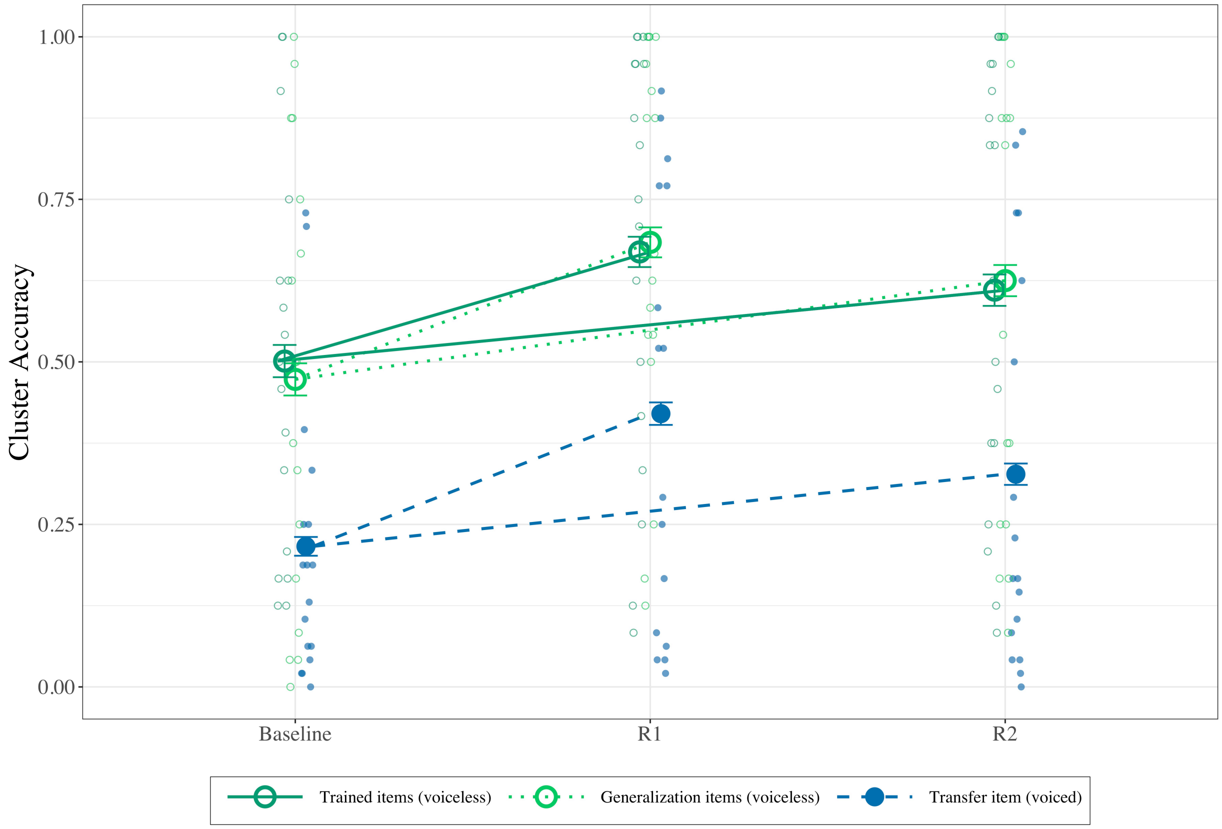

Figure 5. Change in cluster accuracy for the voiceless training condition. The figure depicts overall cluster accuracy for each stimulus group from baseline to R1 and R2. The mean group accuracy was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

Burst-to-burst duration

Figures 6 and 7 present the burst-to-burst duration data from the Voiced training (Figure 6) and Voiceless training (Figure 7) conditions respectively. As can be seen in these figures, there are intrinsic differences in these duration values on voiced clusters and voiceless clusters. In particular, burst-to-burst duration includes the release for the first stop, and that duration will be longer for the voiceless stops than for voiced stops. This leads the burst-to-burst duration to be systematically shorter for voiced clusters than voiceless clusters. In this section, we again consider the results and statistical outcomes relevant to the primary research questions of this paper and revisit this observation in the Discussion.

The best fitting model selected by AIC and BIC was the model that included random intercepts for participant and item, and we also included duration of the stressed vowel following the cluster as discussed above. The model revealed that stressed vowel duration was a significant predictor of burst-to-burst duration overall (β = 64.33, SE = 9.75, $p < .0001$). However, even taking that difference into account, the model revealed significant decreases in burst-to-burst duration from baseline to each retention session for the Trained items (R1: β = -11.38, SE = 1.66, $p < .0001$; R2: β = -11.35, SE = 1.65, $p < .0001$), Generalization items (R1: β = -11.42, SE = 1.67, $p < .0001$; R2: β = -9.71, SE = 1.66, $p < .0001$), and Transfer items (R1: β = -10.66, SE = 1.17, $p < .0001$; R2: β = -7.15, SE = 1.13, $p < .0001$) for the Voiced training condition. In addition, the model indicated that there was a significant difference in the magnitude of change at R2 (β = 4.22, SE = 1.96, $p = 0.03$), where the reduction in duration from baseline for the trained voiced clusters was greater than the reduction for transferred voiceless clusters. No other interaction terms were significant. Overall, these results indicate that participants who practiced voiced clusters produced those trained items with a closer coordination between the two consonants, and that this generalized to untrained nonwords with those clusters, and transferred to the untrained voiceless clusters.

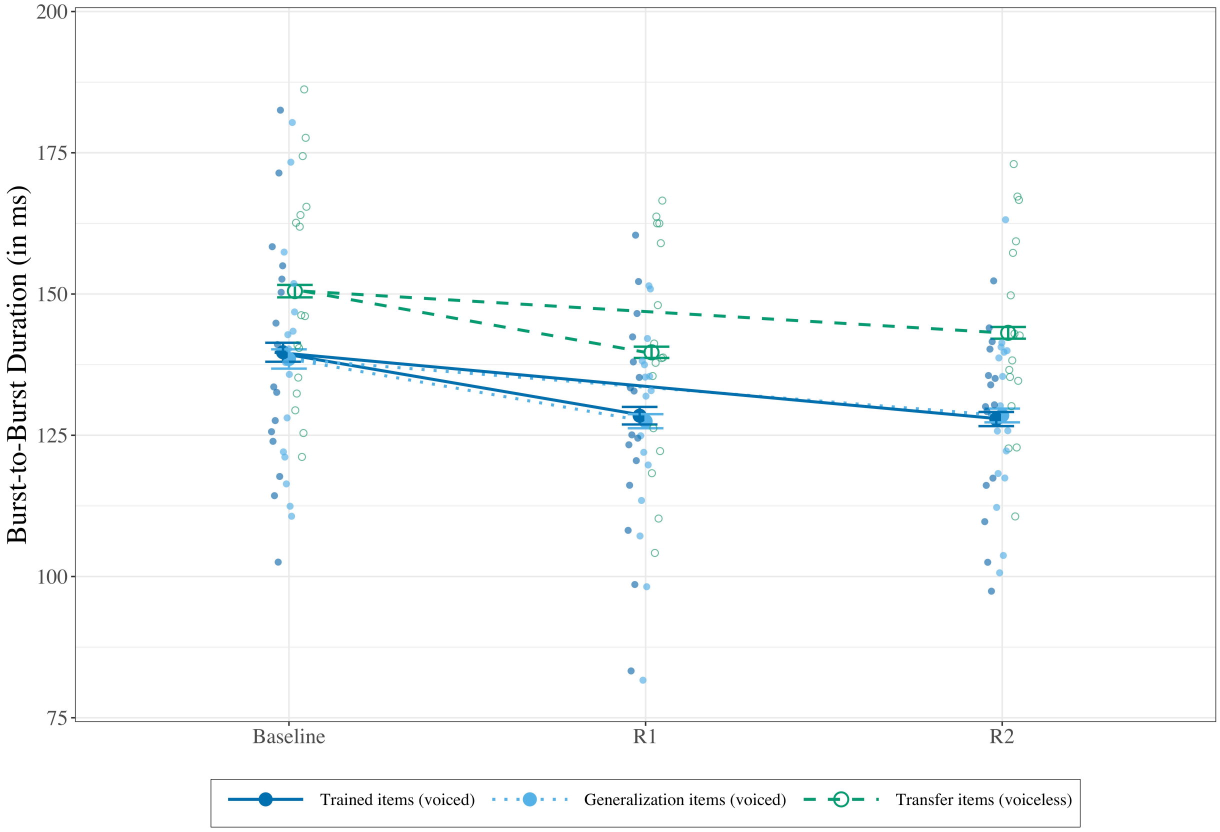

Figure 6. Change in burst-to-burst duration for the voiced training condition. The figure depicts overall burst-to-burst duration for each stimulus group from baseline to R1 and R2. The mean group duration was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

For the Voiceless training condition, the model revealed that there was a significant decrease in burst-to-burst duration from baseline to each retention session for the Trained items (R1: β = -7.06, SE = 1.64, p < .0001; R2: β = -5.6, SE = 1.64, p = .0006), Generalization items (R1: β = -7.68, SE = 1.65, p < .0001; R2: β = -7.12, SE = 1.63, p < .0001), and Transfer items (R1: β = -6.59, SE = 1.17, p < .0001; R2: β = -4.27, SE = 1.17, p = .0003). The interaction between Session and Training was not significant. The results indicate that participants who practiced on voiceless clusters exhibited a decrease in burst-to-burst duration for trained items, and this generalized to untrained voiceless clusters, and transferred to voiced clusters. Thus, although there was a significant interaction between Session and Training for the Voiced training condition but not for the Voiceless training condition, the model did not reveal a significant interaction between Condition, Session and Training. This suggests that the amount of transfer is not asymmetric.

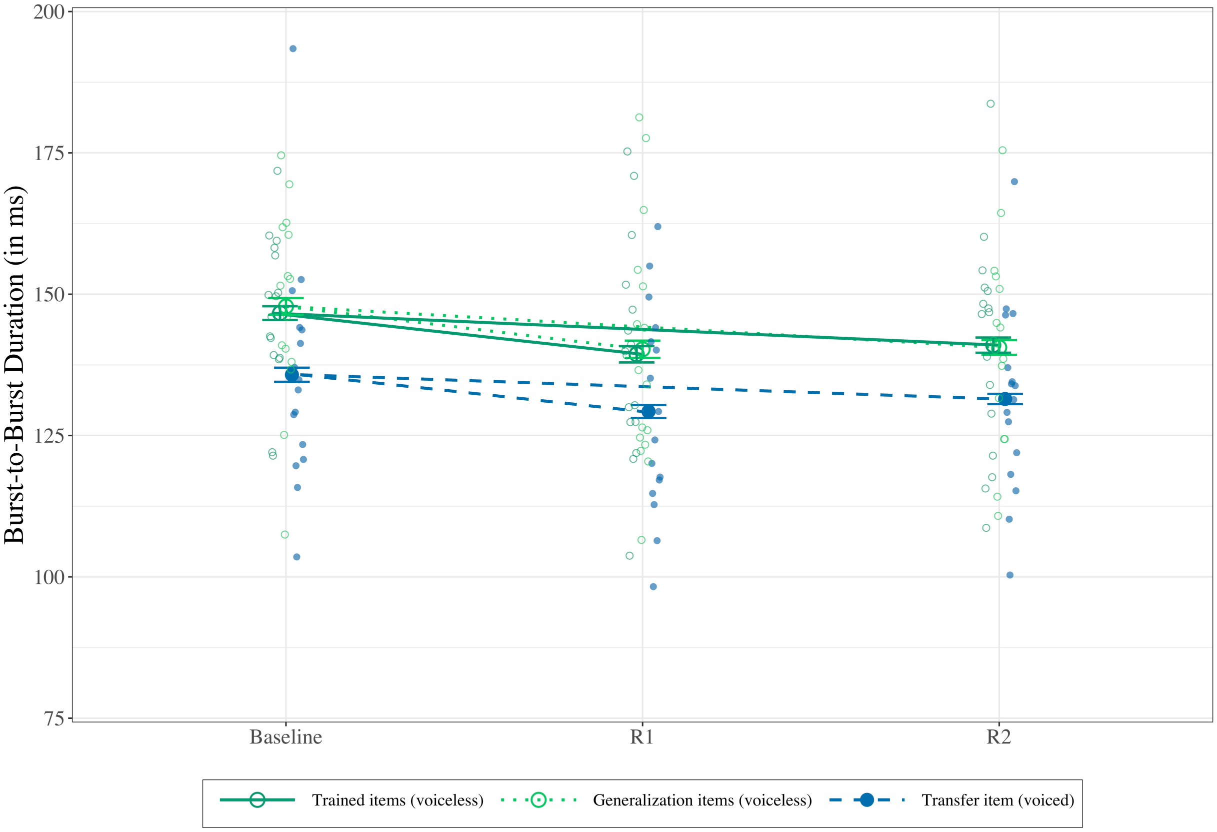

Figure 7. Change in burst-to-burst duration for the voiceless training condition. The figure depicts overall burst-to-burst duration for each stimulus group from baseline to R1 and R2. The mean group duration was plotted against each individual's mean, and the error bars denote

standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

Cluster accuracy: Stop-fricative clusters

Figures 8 and 9 present the cluster accuracy data from the Voiced training (Figure 8) condition and the Voiceless training (Figure 9) condition, respectively. As can be seen in these figures, there was higher accuracy for the voiceless stop-fricative clusters than for the voiced stop-fricative clusters at baseline regardless of the training conditions. This again shows the intrinsic difference in the phonetic implementation between voiced and voiceless stop-fricative clusters. While AIC selected the model that includes random intercepts for both participant and item, BIC selected the model that includes only the random intercept for item. As stated previously, we chose the model that was selected by BIC. The model revealed that, for the Voiced training condition, there was a significant improvement for the voiced stop-fricative clusters from Baseline to both R1 ($\beta = 0.5$, SE = 0.24, $p = .036$) and R2 ($\beta = 0.94$, SE = 0.47, $p = .047$). The same pattern was found for the voiceless stop-fricative clusters, with accuracy improved from Baseline to both R1 ($\beta = 1.18$, SE = 0.2, $p < .0001$) and R2 ($\beta = 0.89$, SE = 0.19, $p < .0001$). In addition, there was a significant difference in the magnitude of change at R1 ($\beta = 0.69$, SE = 0.31, $p = .024$), where the increase in accuracy from baseline for the voiceless stop-fricative clusters was greater than for the voiced stop-fricative clusters. For the Voiceless training condition, there was a significant increase in accuracy for the voiced stop-fricative clusters from Baseline to both R1 ($\beta = 1.43$, SE = 0.23, $p < .0001$) and R2 ($\beta = 0.94$, SE = 0.23, $p < .0001$). Likewise, there was a significant improvement for the voiceless stop-fricative clusters from Baseline to both of the retention sessions (R1: $\beta =1.89$, SE = 0.23, $p < .0001$; R2: $\beta = 1.03$, SE = 0.21, $p < .0001$). There was no significant three-way interaction between Condition, Session, and

Voicing. This suggests that the amount of transfer between the training conditions was not asymmetric. The results revealed that participant also improved on the production of both voiced and voiceless stop-fricative clusters after practicing on either voiced or voiceless stop-stop clusters.

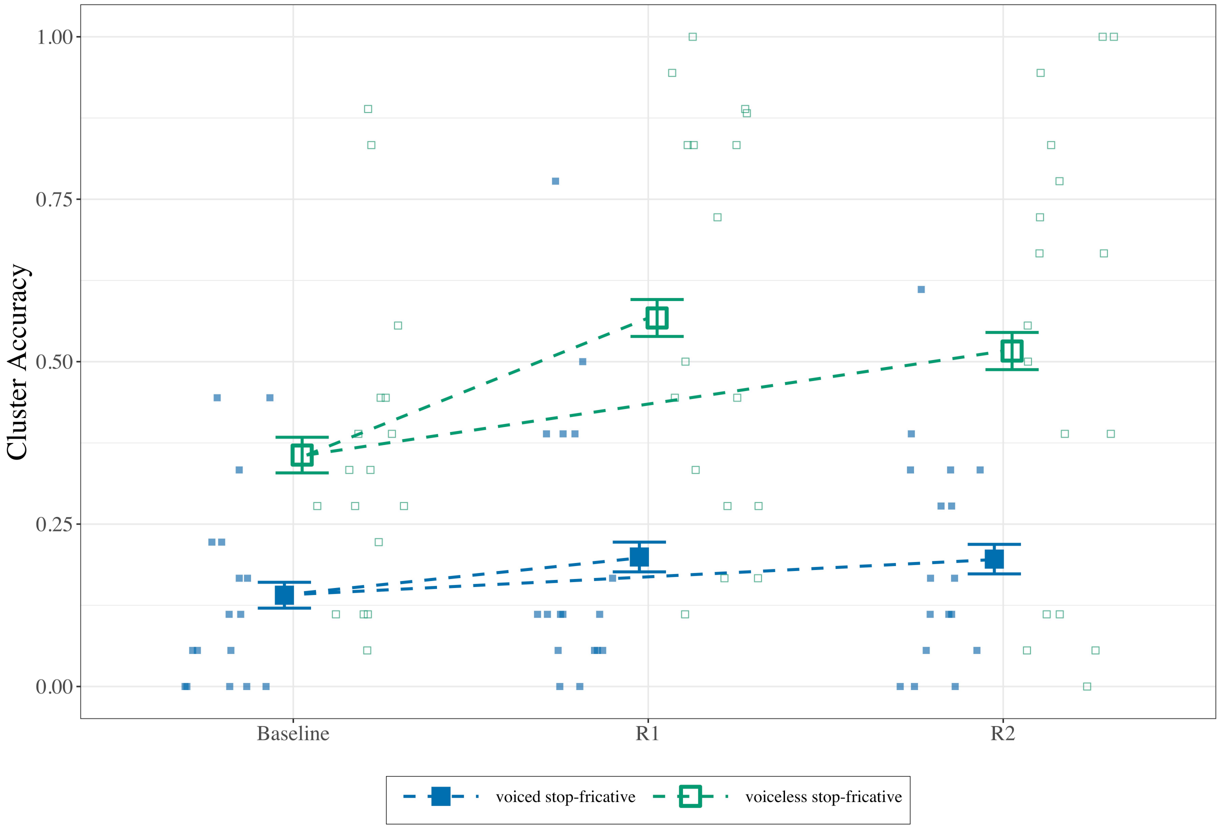

Figure 8. Change in cluster accuracy of stop-fricative clusters for the voiced training condition. The figure depicts overall cluster accuracy for both voiced and voiceless stop-fricative clusters from baseline to R1 and R2. The mean group accuracy was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

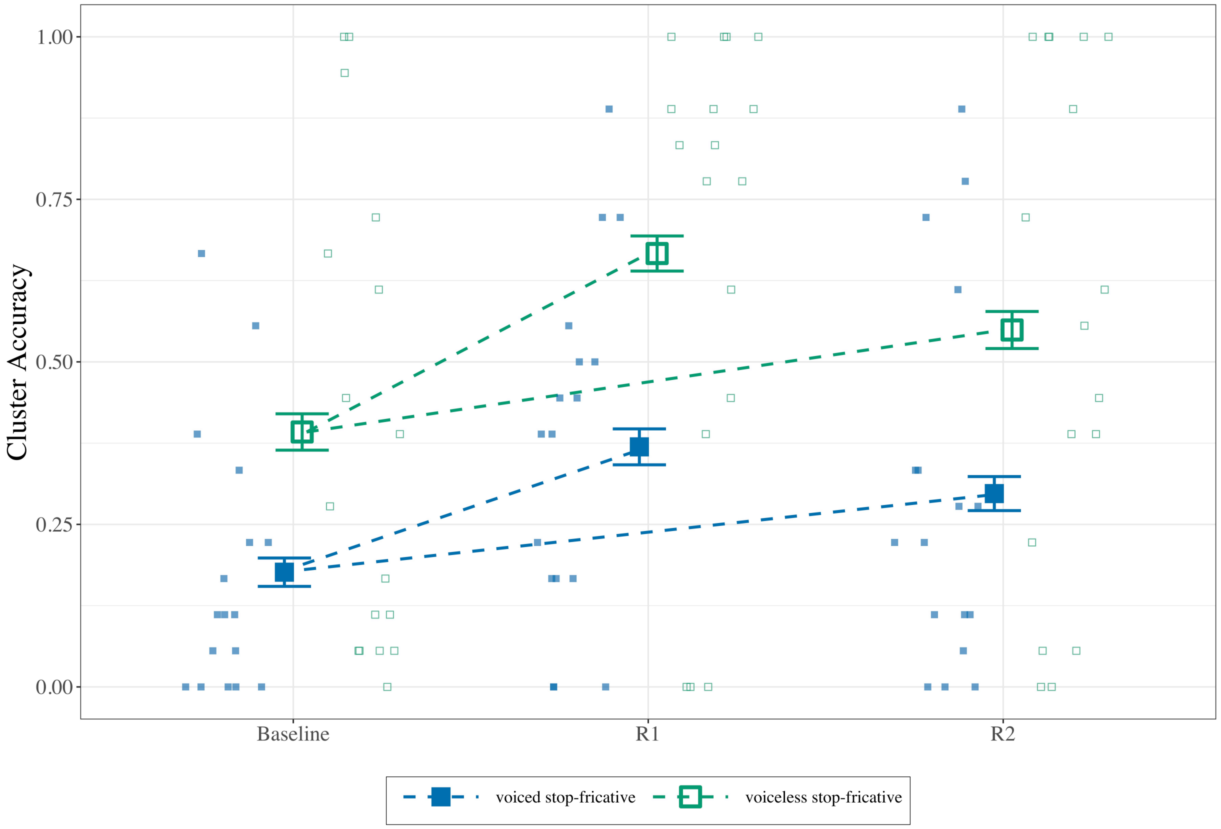

Figure 9. Change in cluster accuracy of stop-fricative clusters for the voiceless training condition. The figure depicts overall cluster accuracy for both voiced and voiceless stop-fricative clusters from baseline to R1 and R2. The mean group accuracy was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

Stop-fricative clusters: C1-C2 duration

Figures 10 and 11 present the C1-C2 duration data for the Voiced training (Figure 10) and the Voiceless training (Figure 11) condition, respectively. As can be seen in these figures, there was a baseline difference in C1-C2 duration between the voiceless stop-fricative clusters and the voiced stop-fricative clusters, regardless of the training conditions. There was a longer C1-C2 duration for the voiceless stop-fricative clusters than the voiced counterparts. This was driven by both voiceless stops having a longer release and voiceless fricative having longer duration. The best-fitting model selected by AIC and BIC was the model that included random intercepts for both participant and item. The model revealed that stressed vowel duration was not a significant predictor of C1-C2 duration. For the Voiced training condition, there was a significant decrease in C1-C2 duration from baseline to both of the retentions for the voiced stop-fricative clusters (R1: $\beta = -12.33$, SE = 2.43, $p < .0001$; R2: $\beta = -10.52$, SE = 2.41, $p < .0001$) and the voiceless stop-fricative clusters (R1: $\beta = -12.08$, SE = 2.33, $p < .0001$; R2: $\beta = -5.7$, SE = 2.32, $p = .014$). For the Voiceless training condition, there was a significant decrease C1-C2 duration from baseline to each retention session for both of the voiced stop-fricative (R1: $\beta = -8.71$, SE = 2.32, $p = .0002$; R2: $\beta = -6.43$, SE = 2.32, $p = .006$) and the voiceless stop-fricative clusters (R1: $\beta = -14.46$, SE = 2.22, $p < .0001$; R2: $\beta = -9.32$, SE = 2.21, $p < .0001$). There was no any significant interaction. Taken together, the data suggests that participants also improved on the coordination for both voiced and voiceless stop-fricative clusters.

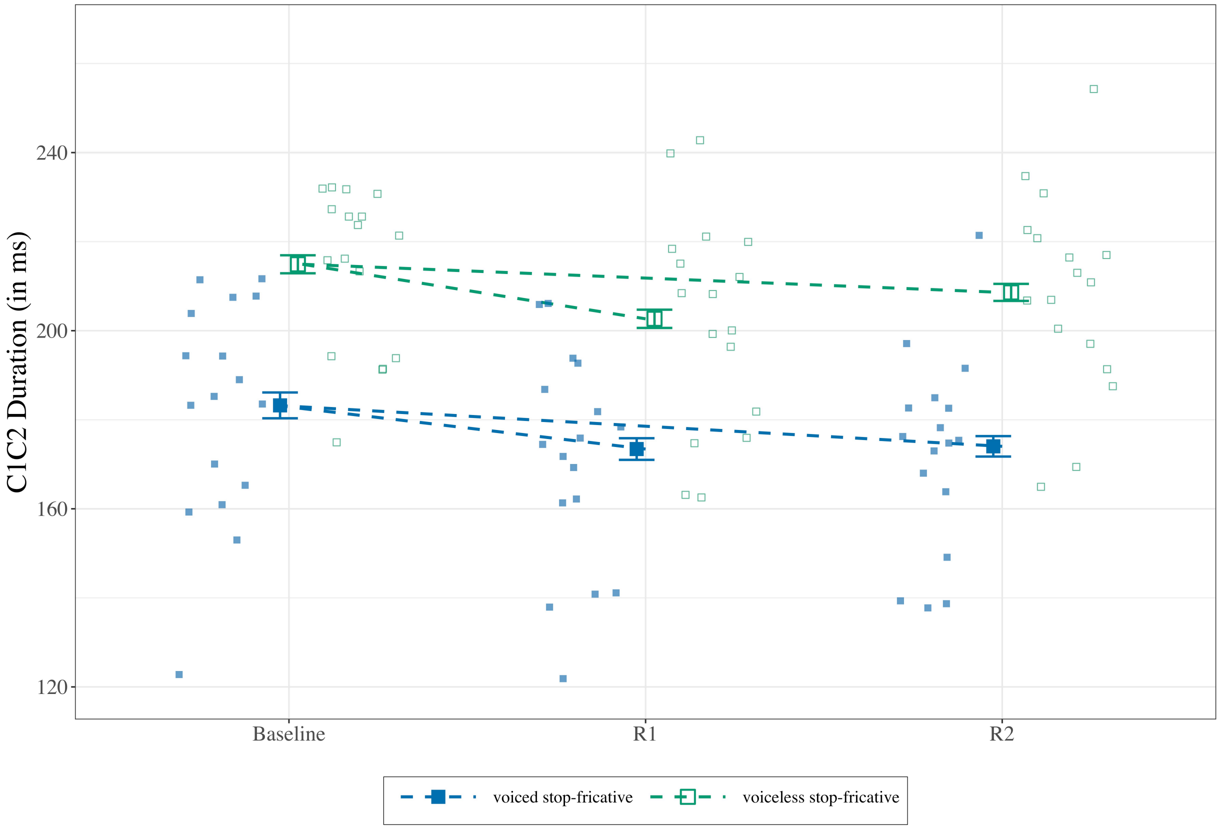

Figure 10. *Change in C1-C2 duration of stop-fricative clusters for the voiced training condition. The figure depicts overall C1-C2 duration for both voiced and voiceless stop-fricative clusters from baseline to R1 and R2. The mean group duration was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.*

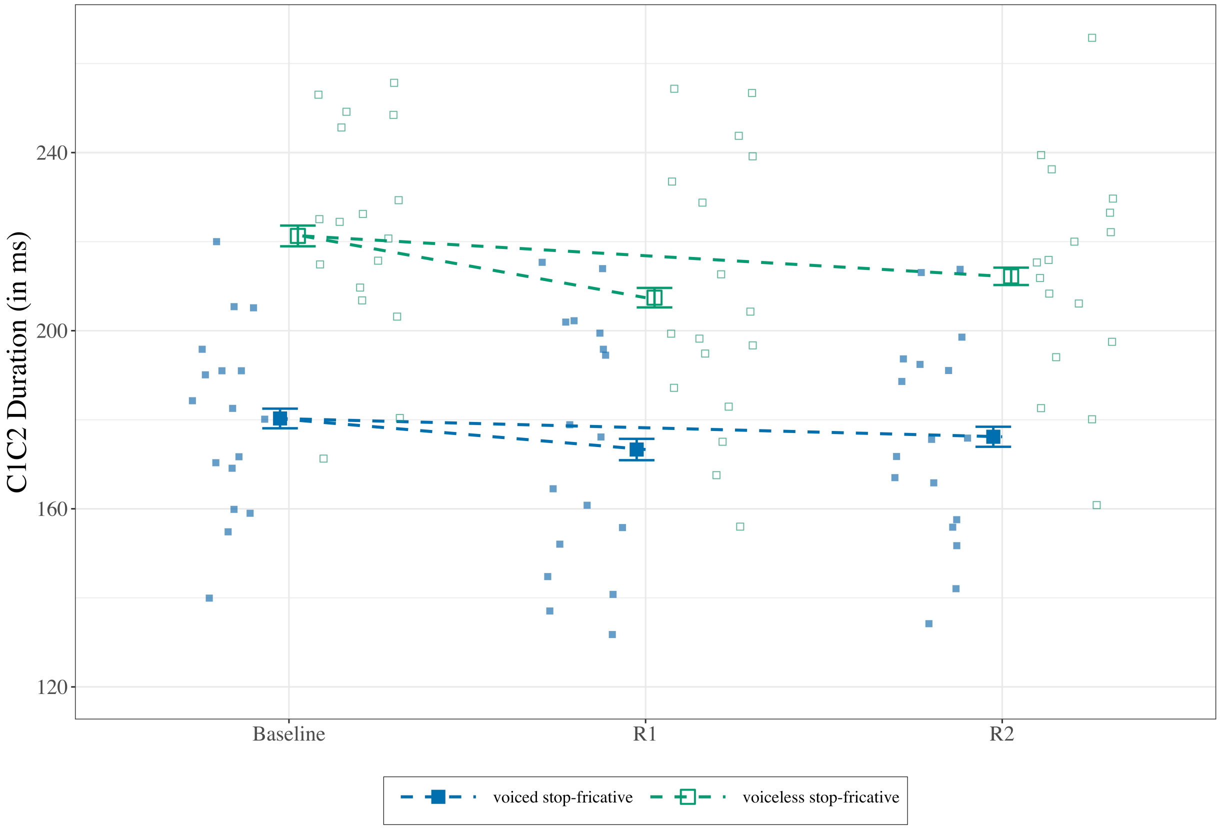

Figure 11. Change in C1-C2 of stop-fricative clusters for the voiceless training condition. The figure depicts overall C1-C2 duration for both voiced and voiceless stop-fricative clusters from baseline to R1 and R2. The mean group duration was plotted against each individual's mean, and the error bars denote standard error. Separate lines connect baseline to R1 and to R2 to reflect our statistical comparison.

Discussion

The current study used a speech motor learning paradigm designed to address three research questions regarding the generalization and transfer of learning in a non-native consonant cluster production task. In particular, we tested the extent to which training on either voiced or voiceless stop-stop clusters leads to improvement on trained items, generalizes to untrained items with the trained clusters, and transfers to the other untrained voicing pattern. Across both accuracy and motor acuity measures, our participants improved on trained items, and generalized to untrained items that contained the trained clusters, as had been previously described in the literature using accuracy and different acoustic measures (Buchwald et al., 2019; Segawa et al., 2019). Moreover, participants in both conditions also improved their accuracy and coordination in producing the clusters from the untrained voicing category.

While the magnitude of improvement between baseline and retention sessions was relatively small, it is worth noting that participants were asked to learn to produce complex speech motor patterns based on a relatively short practice session. The consistent pattern of results suggests that the speech motor learning paradigm was sufficient to facilitate some degree of learning on these complex consonant clusters, and this improvement persisted during the second retention session two days after the practice session. This effect of repetitive practice on learning novel speech motor targets aligned with previous studies (Buchwald et al., 2019; Segawa et al., 2019; Segawa et al., 2015). More importantly, we structured the practice session following the principles of speech motor learning (Maas et al., 2008) (see Methods), including a pre-practice segment to ensure that participants know the targets that they should be attempting during the practice component. The improvement we reported is consistent with the view that these principles can facilitate speech motor learning.

As noted in the results, we found consistent transfer to the untrained cluster type. In addition, post hoc analyses indicated that the participants also improved in their production of stop-fricative clusters following this paradigm. This additional finding raises critical issues about the extent of transfer that we see in speech motor learning tasks, as well as whether the improvement observed in this paradigm is truly an example of motor learning. In the remainder of this section, we discuss how our findings constrain our understanding of the type of non-native onset cluster learning that takes place. We then describe some of the limitations of the present paradigm, and steps to be taken to address these shortcomings in future studies.

Transfer following training on stop-stop clusters

As discussed in the introduction, there exists a limited understanding of how learning novel speech motor sequences transfers to other untrained sequences. Most previous studies have focused on learning at the level of an individual segment, either in the context of acquired speech impairment (Austermann Hula et al., 2008; Ballard et al., 2007; Knock et al., 2000; Wambaugh et al., 1998) or in non-native segment learning (Katz & Mehta, 2015; Li et al., 2019). Our work examined the production of sequences of sounds where the sounds are not novel but their combination in syllable onset is novel. We designed the study to examine whether training on one voicing category of stop-stop clusters would transfer to the other category. Based on the evidence reported here, we believe that speech motor representations encode information about coordination of oral articulators *independently* from information about the coordination of oral and laryngeal articulators. This account would provide an explanation for the fact that learning and transfer within the stop-stop clusters was bidirectional; training on either voiced or voiceless stop-stop clusters led to a significant improvement in the production of the other type of cluster. If information we encode about coordination among articulators did not separate the oral-to-oral coordination from the oral-to-laryngeal coordination, then we would not obtain such a clear result across these conditions.

In designing the experiment, we included a small number of stop-fricative clusters as filler items. Following the main data analysis, we examined the change in performance on these items as well (36 per session vs. 96 per session for the stop-stop clusters) and found an improvement from Baseline to the Retention sessions, both in accuracy and a different motor acuity measure. This post hoc finding showing that the production of stop-fricative clusters also improved requires us to consider our account of transfer more fully. We note two key possible explanations of this finding. The first possibility is that the improvement on the stop-fricative clusters was an additional demonstration of the transfer effect. Under this account, the type of oral-to-oral coordination that was learned during the speech motor learning paradigm would have been sufficient to allow transfer to this other type of sequence. We note that the stop-fricative sequences were designed to be similar to the stop-stop sequences; all were disyllabic nonwords with a 'back-to-front' coordination pattern (i.e., the first consonant had a more posterior place of constriction than the second consonant). We also note that there is evidence that stop-stop clusters are more complex than stop-fricative clusters, both with respect to the more limited cross-linguistic distribution of stop-stop clusters (Morelli, 1999) and their baseline accuracy (Davidson, 2010). While previous speech motor learning studies had not reported transfer across manner of articulation (e.g., Ballard et al., 2007), those studies examined singletons which have different articulatory mechanisms from the consonant clusters examined here. Given these factors, we believe that it is likely that the improvement of stop-fricative items reflects an additional example of transfer of learning, although we also believe that this can be addressed empirically in future work as outlined below.

An alternative account of this improvement is that the practice component of the speech motor learning paradigm was not critical to the improvement, and that the improvement seen across clusters derived from the straightforward instruction in the pre-practice session for how to produce a consonant cluster. With respect to this account, we note that this pre-practice session focused on different cluster types than those tested in this study, as the pre-practice focused on fricative-stop and fricative-nasal. We believe that this instruction is likely to be necessary to promote learning of these complex non-native consonant clusters, as pre-practice is a critical component of the motor learning paradigm and has been used in previous studies of non-native cluster learning (Buchwald et al., 2019; Segawa et al., 2019). However, it is not clear whether this instruction is sufficient to lead to the widespread improvement we observed. If this instruction were indeed the locus of the improvement, and not the practice session, then we do not believe that these findings would actually reflect motor learning. In a previous study that did not include a separate baseline session, Buchwald et al. (2019) examined performance throughout the practice session and found improvement from the beginning to the end, suggesting that the practice is critical to learning. However, to rule out the possibility that the instruction alone can drive this type of systematic improvement, we will need to run a different experimental condition in which participants receive that same instruction but then do not practice non-native consonant clusters during the practice session. If the improvement across these difficult clusters is still observed, we would then be forced to conclude that the practice is not the cause of the improvement. However, if the improvement is not seen in the absence of practice, then we must conclude that the practice is also crucial to the cluster learning.

Effect of complexity on transfer

In the introduction, we argued that if complexity of the targets affected the transfer, this would lead to an asymmetry, with more transfer from voiced to voiceless stop-stop clusters than the other direction. We did not find support for this in our data. We consider here that our definition of complexity did not actually reflect the specific differences in terms of the complexity of learning to produce these clusters, even though this difference is supported by phonetic and phonological evidence discussed in the introduction. We did find consistently large differences in terms of cluster accuracy, with voiceless clusters more accurate at all stages of the study as has been observed in other studies (Davidson, 2006, 2010). However, it is possible that this accuracy difference was partly an artifact of our analysis, as epenthesis may be harder to observe in the acoustics in voiceless stop-stop clusters. We observed a large number of vowel epenthesis errors in the voiced stop-stop clusters; however, a speaker may have the same oral articulator coordination in producing a voiceless stop-stop cluster, but an absence of phonation would lead this to be unobservable on the acoustic record. We note that we still observed improvement in both cluster types, so it is likely that something was being learned and modified by these speakers. However, it remains possible that the aspect of these stop-stop clusters that is particularly difficult for speakers to learn to produce is unrelated to the inherent differences between these clusters.

In the previous section, we argued that it is likely that the improvement we observed on stop-fricative clusters may be attributable to transfer of learning. We also noted that stop-stop clusters are considered more complex than stop-fricative clusters. To follow-up on the complexity issue as well as the issue transfer issue discussed above, we plan to run an additional study in which we train participants on stop-fricative clusters and then test them on both stop-stop and stop-fricative clusters. This will allow us to explore the complexity issue within the oral-to-oral articulator patterns alone. However, if the observed improvement and transfer was driven solely by the pre-practice instruction alone, as considered above, then we would not expect any effect of the complexity of the trained items the magnitude of transfer. Again, this possibility requires further examination when the aforementioned control groups are included.

Limitations and future directions

Within the scope of the original research questions, the present findings demonstrated a bidirectional transfer pattern between voicing categories; however, our design did not permit us to address whether there was transfer within the trained voicing category (e.g., from trained voiced clusters to untrained and different voiced clusters). Further work is needed to address this question. For example, by including stop-stop clusters with untrained front-to-back articulation (e.g., /bd/, or /tk/), we could examine whether there is transfer to clusters with same voicing pattern but untrained oral-to-oral articulator transition. Another potential direction is to manipulate the vowel context following the onset clusters. Given that we consistently used /i/, /ɑ/, /u/ as the nucleus in the first syllable in both trained and untrained items, a future study could include a different vowel that is not practiced. Adding this manipulation would allow us to test whether the learning can transfer to a different vowel context.

In addition, we discussed above how our reliance on the acoustic record may have artificially deflated the number of vowel insertion errors observed in the voiceless stop-stop clusters. We do not believe that this drove any crucial effects; this limitation may have affected the analysis of all voiceless clusters, but we still observed a clear and consistent improvement in these sequences. However, it will be important to continue to examine these coordination issues using articulatory measures such as electromagnetic articulography (EMA).

Finally, as we learned through our post hoc analyses, in order to ask questions about the specificity of speech motor representations, it will be critical to include a complete control condition in the future containing items that we do not expect to improve. This will allow us to more completely address the nature and content of speech motor representations.

Conclusion

The present study used a practice-based speech motor learning paradigm to investigate the transfer patterns following training on either voiced or voiceless stop-stop clusters. Our data show that participants improved on the trained clusters in both trained and untrained stimuli, and also improved in their production of the untrained cluster type. We argue that this pattern of transfer arises because the temporal coordination of oral-to-oral articulators is encoded independently from that of oral-to-laryngeal articulators. In a post hoc analysis, we further observed widespread improvement on stop-fricative clusters originally included only as filler items which we interpret here as an additional transfer effect, although additional work will be needed to rule out alternative explanations. Future studies are needed to further investigate the specificity of learned speech motor representations in non-native clusters and to shed light on the underlying mechanism of practice-based speech motor learning paradigm.

Appendix A1 : International phonetic alphabet (IPA) transcription and orthography for target stimuli

Target stimuli

Cluster	IPA	Orthography	IPA	Orthography
/gd/	[gdimu]	GDEEMOO	[gdabi]	GDAHBEE
	[gdanæd]	GDAHNAD	[gduzæb]	GDOOZAB
	[gdubmɑt]	GDOOBMOT	[gdinbud]	GDEENBOOD
	[gdikpræd]	GDEEKPRAD	[gdumprid]	GDOOMPREED
/gb/	[gbimu]	GBEEMOO	[gbafu]	GBAHFOO
	[gbɑdæst]	GBAHDAST	[gbudæp]	GBOODAP
	[gbumdut]	GBOOMDOOT	[gbinzɑm]	GBEENZOM
	[gbinflɑt]	GBEENFLOT	[gbultræp]	GBOOLTRAP
/db/	[dbagi]	DBAHGEE	[dbidu]	DBEEDOO
	[dbudæp]	DBOODAP	[dbamæk]	DBAHMAK
	[dbigzun]	DBEEGZOON	[dbugbɑt]	DBOOGBOT
	[dbutgrin]	DBOOTGREEN	[dbitflæg]	DBEETFLAH
/kt/	[ktigu]	KTEEGOO	[ktɑni]	KTAHNEE
	[ktɑmæk]	KTAHMACK	[ktupæb]	KTOOPAB
	[ktubʃɑp]	KTOOBSHOP	[ktibgun]	KTEEBGOON
	[ktɑksnæm]	KTAHKSNAM	[ktudsmik]	KTOODSMEEK
/kp/	[kpibu]	KPEEBOO	[kpazi]	KPAHZEE
	[kpɑdæm]	KPAHDAM	[kpugæn]	KPOOGAN
	[kpuʃpɑk]	KPOOSHPOK	[kpitmuk]	KPEETMOOK
	[kpakspæd]	KPAHSHPAD	[kpugdwim]	KPOOGDWEEM
/tp/	[tpɑdi]	TPAHDEE	[tpidu]	TPEEDOO
	[tpudæf]	TPOODAF	[tpɑgæm]	TPAHGAM
	[tpɑmgut]	TPAHMGOOT	[tputgɑb]	TPOOTGOB
	[tpidprɑb]	TPEEDPROB	[tpɑbtræn]	TPAHBTRAN

Appendix A2 : International phonetic alphabet (IPA) transcription and orthography for filler stimuli

Filler stimuli in the baseline, R1, and R2 phase

Cluster	IPA	Orthography	Cluster	IPA	Orthography
/gv/	[gvɑni]	GVAHNEE	/kf/	[kfɑdi]	KFAHDEE
	[gvidbræm]	GVEEDBRAM		[kfudæb]	KFOODAB
	[gvudmɑk]	GVOODMOCK		[kfidblum]	KFEEDBLOOM
/gz/	[gzɑdæf]	GZAHDAF	/ks/	[ksɑbi]	KSAHBEE
	[gzidu]	GZEEDOO		[ksukbɑm]	KSOOKBOM
	[gzudbrit]	GZOODBREET		[ksidzud]	KSEEDZOOD
/dv/	[dvɑgæp]	DVAHGAP	/tf/	[tfɑsæb]	TFAHSAB
	[dvigu]	DVEEGOO		[tfidu]	TFEEDOO
	[dvutʃrig]	DVOOTSHREEG		[tfukswig]	TFOOKSWEEG
/fl/	[flɑpstæn]	FLAHPSTAN	/sn/	[snɑmi]	SNAHMEE
	[flinæd]	FLEENAD		[snidtwæg]	SNEEDTWAG
	[fluvi]	FLOOVEE		[snuzæn]	SNOOZAN
/sl/	[slɑdi]	SLAHDEE			
	[slikbrit]	SLEEKBREET			
	[sludæm]	SLOODAM			

Filler stimuli in the practice phase

Cluster	IPA	Orthography	Singleton	IPA	Orthography
/bl/	[blugɑ]	BLOOGAH	/l/	[ligu]	LEEGOO
	[bliwæn]	BLEEWAN		[lɑdæp]	LAHDAP
/fr/	[frutswin]	FROOTSWEEN	/r/	[rugæn]	ROOGAN
	[frɑvæp]	FRAHVAP		[rɑvi]	RAHVEE
/sm/	[smidu]	SMEEDOO	/w/	[winu]	WEENOO
	[smutflæm]	SMOOTFLAM		[wubɑm]	WOOBOM

Acknowledgments

This work was supported by a grant awarded from National Institute on Deafness and other Communication Disorders (NIDCD) to Adam Buchwald (K01DC014298). The authors would like to thank Maria Grigos and Tara McAllister for their comments and inputs on this study; Megan Burns, Alexandra Gordon, Izabela Grzebyk, Kevin Tjokro, and Yulia White for their help on data collection and analysis.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Austermann Hula, S. N., Robin, D. A., Maas, E., Ballard, K. J., & Schmidt, R. A. (2008). Effects of feedback frequency and timing on acquisition, retention, and transfer of speech skills in acquired apraxia of speech. *J Speech, Language, and Hearing Resarch*, *51*(5), 1088-1113. https://doi.org/10.1044/1092-4388(2008/06-0042)

Ballard, K. J. (2001). Response generalization in apraxia of speech treatments: taking another look. *Journal of Communication Disorders*, *34*(1-2), 3-20.

Ballard, K. J., Maas, E., & Robin, D. A. (2007). Treating control of voicing in apraxia of speech with variable practice. *Aphasiology*, *21*(12), 1195-1217. https://doi.org/10.1080/02687030601047858

Bates, D., Maechler, M., & Bolker, B. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer [Computer program].* In (Version 6.1.03) http://www.praat.org/

Bombien, L., & Hoole, P. (2013). Articulatory overlap as a function of voicing in French and German consonant clusters. *The Journal of the Acoustical Society of America*, *134*(1), 539-550. https://doi.org/10.1121/1.4807510

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, *45*(2-4), 140-155. https://doi.org/10.1159/000261823

Browman, C. P., & Goldstein, L. (1995). Dynamics and Articulatory Phonology. In *Mind as motion*.

Buchwald, A., Calhoun, H., Rimikis, S., Lowe, M. S., Wellner, R., & Edwards, D. J. (2019). Using tDCS to facilitate motor learning in speech production: The role of timing. *Cortex*, *111*, 274-285. https://doi.org/10.1016/j.cortex.2018.11.014

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, *24*, 209-244.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. *The Journal of the Acoustical Society of America*, *128*(4), 2033-2048.

Carey, D., Miquel, M. E., Evans, B. G., Adank, P., & McGettigan, C. (2017). Functional brain outcomes of L2 speech learning emerge during sensorimotor transformation. *Neuroimage*, *159*, 18-31. https://doi.org/10.1016/j.neuroimage.2017.06.053

Caudrelier, T., Schwartz, J. L., Perrier, P., Gerber, S., & Rochet-Capellan, A. (2018). Transfer of Learning: What Does It Tell Us About Speech Production Units? *Journal of Speech, Language, and Hearing Research*, *61*(7), 1613-1625. https://doi.org/10.1044/2018_JSLHR-S-17-0130

Chitoran, I., Goldstein, L., & Byrd, D. (2002). Gestural overlap and recoverability: Articulatory evidence from Georgian. *Papers in Laboratory Phonology*, *7*, 419-447.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, *34*(1), 104-137. https://doi.org/10.1016/j.wocn.2005.03.004

Davidson, L. (2007). The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology*, *24*(2), 261-286. https://doi.org/10.1017/s0952675707001200

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics*, *38*(2), 272-288. https://doi.org/10.1016/j.wocn.2010.01.001

Davidson, L., & Roon, K. (2008). Durational correlates for differentiating consonant sequences in Russian. *Journal of the International Phonetic Association*, *38*(2), 137-165. https://doi.org/10.1017/s0025100308003447

Davis, M. J. (2010). Contrast coding in multiple regression analysis: Strengths, weaknesses, and utility of popular coding structures. *Journal of Data Science*, *8*(1), 61-73.

Harel, D., & McAllister, T. (2019). Multilevel Models for Communication Sciences and Disorders. *Journal of Speech, Language, and Hearing Research*, *62*(4), 783-801. https://doi.org/10.1044/2018_JSLHR-S-18-0075

Hoole, P., & Bombien, L. (2014). Laryngeal–oral coordination in mixed-voicing clusters. *Journal of Phonetics*, *44*, 8-24. https://doi.org/10.1016/j.wocn.2014.02.004

Hoole, P., & Bombien, L. (2017). A Cross-Language Study of Laryngeal-Oral Coordination Across Varying Prosodic and Syllable-Structure Conditions. *Journal of Speech, Language, and Hearing Research*, *60*(3), 525-539. https://doi.org/10.1044/2016_JSLHR-S-15-0034

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213-1216. https://doi.org/10.1126/science.279.5354.1213

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817-832. https://doi.org/10.1121/1.4926561

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & 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. https://doi.org/10.1016/j.wocn.2016.05.001

Kartushina, N., & Martin, C. D. (2019). Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training. *Language Learning*, *69*(1), 71-105. https://doi.org/10.1111/lang.12315

Katz, W. F., & Mehta, S. (2015). Visual Feedback of Tongue Movement for Novel Speech Sound Learning. *Frontiers in Human Neuroscience*, *9*, 612. https://doi.org/10.3389/fnhum.2015.00612

Kawasaki-Fukumori, H., & Ohala, J. (1997). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In S. Eliasson & E. H. Jahr (Eds.), *Language and its ecology: Essays in memory of Einar Haugen* (pp. 343).

Knock, T. R., Ballard, K. J., Robin, D. A., & Schmidt, R. A. (2000). Influence of order of stimulus presentation on speech motor learning: A principled approach to treatment for apraxia of speech. *Aphasiology*, *14*(5-6), 653-668. https://doi.org/10.1080/026870300401379

Levitt, J. S., & Katz, W. F. (2007). Augmented visual feedback in second language learning: Training Japanese post-alveolar flaps to American English speakers.

Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America*, *146*(6), 4625. https://doi.org/10.1121/1.5139423

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics*, *8*, 475-490.

Löfqvist, A., & Yoshioka, H. (1980). Laryngeal activity in Swedish obstruent clusters. *The Journal of the Acoustical Society of America*, *68*(3), 792-801. https://doi.org/10.1121/1.384774

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication*, *3*(4), 279-289.

Maas, E., Barlow, J., Robin, D., & Shapiro, L. (2002). Treatment of sound errors in aphasia and apraxia of speech: Effects of phonological complexity. *Aphasiology*, *16*(4-6), 609-622. https://doi.org/10.1080/02687030244000266

Maas, E., Robin, D. A., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of Motor Learning in Treatment of Motor Speech Disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277-298. https://doi.org/10.1044/1058-0360(2008/025)

Marin, S., & Pouplier, M. (2010). Temporal Organization of Complex Onsets and Codas in American English: Testing the Predictions of a Gestural Coupling Model. *Motor Control*, *14*(3), 380-407. https://doi.org/DOI 10.1123/mcj.14.3.380

Morelli, F. (1999). *The phonotactics and phonology of obstruent clusters in Optimality Theory* University of Maryland].

Ohala, J. (1983). The Origin of Sound Patterns in Vocal Tract Constraints. In *The production of speech* (pp. 189-216). Springer.

Ohala, J. (1997). Aerodynamics of phonology. Seoul International Conference on Linguistics,

Pastätter, M., & Pouplier, M. (2017). Articulatory mechanisms underlying onset-vowel organization. *Journal of Phonetics*, *65*, 1-14. https://doi.org/10.1016/j.wocn.2017.03.005

Pierece, J. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.

Pouplier, M., Marin, S., Hoole, P., & Kochetov, A. (2017). Speech rate effects in Russian onset clusters are modulated by frequency, but not auditory cue robustness. *Journal of Phonetics*, *64*, 108-126. https://doi.org/10.1016/j.wocn.2017.01.006

R Core Team. (2017). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing.

Repp, B. H., & Lin, H. B. (1989). Acoustic properties and perception of stop consonant release transients. *J Acoust Soc Am*, *85*(1), 379-396. https://doi.org/10.1121/1.397689

Riecker, A., Brendel, B., Ziegler, W., Erb, M., & Ackermann, H. (2008). The influence of syllable onset complexity and syllable frequency on speech motor control. *Brain & Language*, *107*(2), 102-113. https://doi.org/10.1016/j.bandl.2008.01.008

Rochet-Capellan, A., Richer, L., & Ostry, D. J. (2012). Nonhomogeneous transfer reveals specificity in speech motor learning. *Journal of Neurophysiology*, *107*(6), 1711-1717. https://doi.org/10.1152/jn.00773.2011

Sadagopan, N., & Smith, A. (2008). Developmental changes in the effects of utterance length and complexity on speech movement variability. *Journal of Speech, Language, and Hearing Research*, *51*(5), 1138-1151. https://doi.org/10.1044/1092-4388(2008/06-0222)

Schneider, S., & Frens, R. (2005). Training four-syllable CV patterns in individuals with acquired apraxia of speech: Theoretical implications. *Aphasiology*, *19*(3-5), 451-471. https://doi.org/10.1080/02687030444000886

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 464.

Segawa, J., Masapollo, M., Tong, M., Smith, D. J., & Guenther, F. H. (2019). Chunking of phonological units in speech sequencing. *Brain & Language*, *195*, 104636. https://doi.org/10.1016/j.bandl.2019.05.001

Segawa, J., Tourville, J. A., Beal, D. S., & Guenther, F. H. (2015). The neural correlates of speech motor sequence learning. *Journal of Cognitive Neuroscience*, *27*(4), 819-831. https://doi.org/10.1162/jocn_a_00737

Steinberg Lowe, M., & Buchwald, A. (2017). The Impact of Feedback Frequency on Performance in a Novel Speech Motor Learning Task. *Journal of Speech, Language, and Hearing Research*, *60*(6S), 1712-1725. https://doi.org/10.1044/2017_JSLHR-S-16-0207

Thompson, C. K., Shapiro, L. P., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia: the complexity account of treatment efficacy (CATE). *Journal of Speech, Language, and Hearing Research*, *46*(3), 591-607. https://doi.org/10.1044/1092-4388(2003/047)

Tremblay, S., Houle, G., & Ostry, D. J. (2008). Specificity of speech motor learning. *The Journal of Neuroscience*, *28*(10), 2426-2434. https://doi.org/10.1523/JNEUROSCI.4196-07.2008

Wambaugh, J. L., West, J. E., & Doyle, P. J. (1998). Treatment for apraxia of speech: Effects of targeting sound groups. *Aphasiology*, *12*(7-8), 731-743. https://doi.org/10.1080/02687039808249569

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. In https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. In (Version R package version 0.8.1) https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *tidyr: Easily Tidy Data with "spread()" and "gather()" Functions*. In (Version R package version 0.8.3) https://CRAN.R-project.org/package=tidyr

Wilson, C., Davidson, L., & Martin, S. (2014). Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language*, *77*, 1-24. https://doi.org/10.1016/j.jml.2014.08.001