Individual predictors of response to biofeedback training for second-language production

Joanne Jingwen Li,[1] Samantha Ayala,[1] Daphna Harel,[2] Douglas M. Shiller,[3] and Tara McAllister[1,a)]

[1]*Department of Communicative Sciences and Disorders, New York University, 665 Broadway, Suite 900, New York, New York 10012, USA*

[2]*Department of Applied Statistics, Social Science, and Humanities, New York University, 246 Greene Street, 3rd Floor, New York, New York 10003, USA*

[3]*École d'orthophonie et d'audiologie, Université de Montréal, Case Postale 6128, Succursale Centre-ville, Montréal, Québec, H3C 3J7, Canada*

(Received 7 April 2019; revised 14 November 2019; accepted 14 November 2019; published online 30 December 2019)

While recent research suggests that visual biofeedback can facilitate speech production training in clinical populations and second language (L2) learners, individual learners' responsiveness to biofeedback is highly variable. This study investigated the hypothesis that the type of biofeedback provided, visual-acoustic versus ultrasound, could interact with individuals' acuity in auditory and somatosensory domains. Specifically, it was hypothesized that learners with lower acuity in a sensory domain would show greater learning in response to biofeedback targeting that domain. Production variability and phonological awareness were also investigated as predictors. Sixty female native speakers of English received 30 min of training, randomly assigned to feature visual-acoustic or ultrasound biofeedback, for each of two Mandarin vowels. On average, participants showed a moderate magnitude of improvement (decrease in Euclidean distance from a native-speaker target) across both vowels and biofeedback conditions. The hypothesis of an interaction between sensory acuity and biofeedback type was not supported, but phonological awareness and production variability were predictive of learning gains, consistent with previous research. Specifically, high phonological awareness and low production variability post-training were associated with better outcomes, although these effects were mediated by vowel target. This line of research could have implications for personalized learning in both L2 pedagogy and clinical practice. © *2019 Acoustical Society of America.*
https://doi.org/10.1121/1.5139423

[BVT] Pages: 4625–4643

I. INTRODUCTION

A. Biofeedback in speech learning

Learning to produce speech sounds accurately in a second language (L2) can be challenging, and most speakers who acquire an L2 after childhood show persisting production differences that are described as a foreign accent. Previous research has explored different training paradigms in an effort to understand how successful acquisition of pronunciation of an L2 can best be achieved (e.g., Akahane-Yamada *et al.*, 1998; Catford and Pisoni, 1970; Engwall, 2006; Wong, 2013). However, L2 production remains an under-studied area (e.g., Gilbert, 2010), with an ongoing need for further research to understand both what training paradigms are most successful in yielding native-like pronunciation, and what individual-level factors can best account for variability in outcomes in L2 pronunciation learning.

Various studies have suggested that visual biofeedback may facilitate speech sound learning in individuals with speech disorders (e.g., Adler-Bock *et al.*, 2007; Hardcastle *et al.*, 1991; McAllister Byun, 2017; McAllister Byun and Hitchcock, 2012; McAllister Byun *et al.*, 2014; Preston and Leaman, 2014) as well as L2 learners (Gick *et al.*, 2008; Kartushina *et al.*, 2015). Biofeedback involves the use of instrumentation to generate a real-time visual display of speech. The learner can observe this display and alter their output in an effort to achieve a better match with a model representing correct production. There are multiple technologies used for visual biofeedback that differ in the information provided. For example, visual-acoustic biofeedback provides a visualization of the spectrum of the acoustic signal of speech (e.g., McAllister Byun and Hitchcock, 2012), ultrasound biofeedback provides a real-time view of the shape and movement of the tongue inside the oral cavity (e.g., Adler-Bock *et al.*, 2007), and electropalatography (EPG) shows the placement and timing of contact between the tongue and the hard palate (e.g., Hardcastle *et al.*, 1991; Hitchcock *et al.*, 2017). The present study focuses on visual-acoustic and ultrasound biofeedback, illustrated in Figs. 1 and 2, in the context of an L2 speech sound training task.

In a meta-analysis that surveyed the literature on computer-assisted visual feedback for L2 pronunciation training, Bliss *et al.* (2018) argued that technology can have a positive impact on pronunciation training by providing information that is less accessible through conventional channels. For example, ultrasound biofeedback provides explicit information about movements of the tongue that would ordinarily be

a)Electronic mail: tkm214@nyu.edu

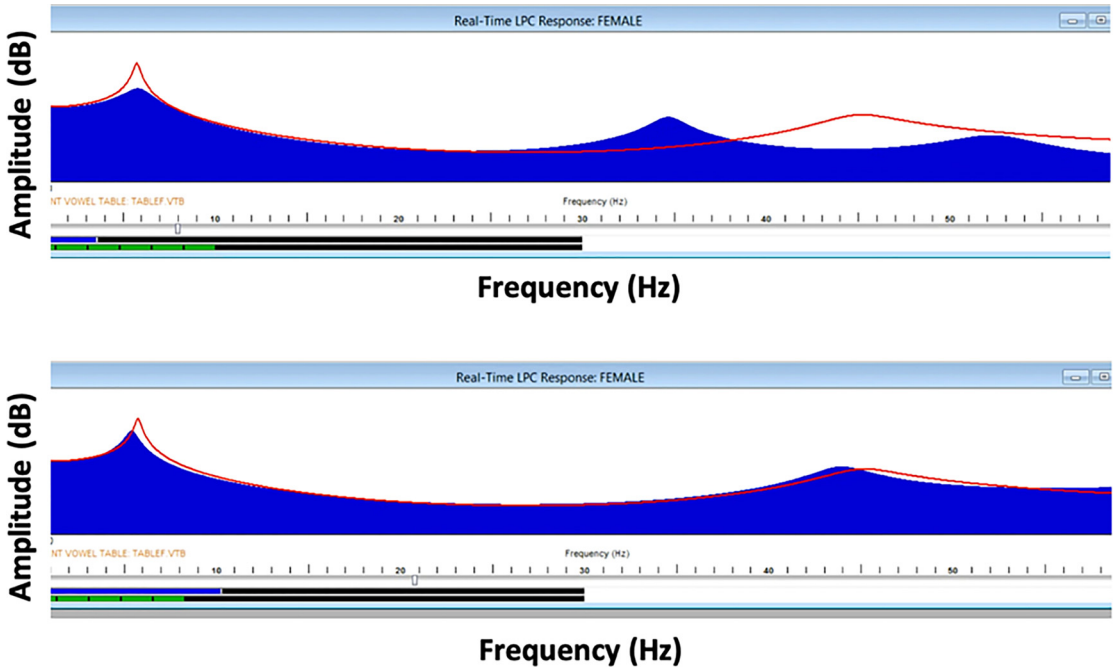

FIG. 1. (Color online) Examples of visual-acoustic biofeedback displays generated with KayPentax Sona-Match software (PENTAX MEDICAL). Incorrect (top) and correct (bottom) productions of the vowel /y/ contrast in the frequency location of the second formant peak in a Linear Predictive Coding (LPC) spectrum.

visually concealed within the oral cavity. Gick *et al.* (2008) reported that Japanese learners of English as an L2 showed a significant improvement in their production of the English /ɹ/-/l/ contrast, which is typically highly challenging, after 30 min of ultrasound feedback training. Positive effects on L2 pronunciation have also been reported when training is enhanced with visual-acoustic biofeedback. Kartushina *et al.* (2015) trained first language (L1) French speakers to produce Danish vowels that were not part of their native vowel inventory. Participants who received one hour of visual-acoustic biofeedback training per vowel demonstrated significant improvement in their productions, whereas a non-biofeedback control group demonstrated no significant improvement despite hearing and producing the same number of repetitions of the Danish vowels. Additional published reports of positive effects of biofeedback training on L2 production outcomes include Dowd *et al.* (1998), Carey (2004), Kocjančičo Antolík *et al.* (2019), and Kartushina *et al.* (2016) for spectral characteristics of vowels; Okuno and Hardison (2016) for vowel length contrasts; and Olson (2014) for fricative production. Although well-controlled comparisons of training with and without biofeedback remain sparse, studies like the above show promise for the use of biofeedback technologies in L2 pronunciation training.

Biofeedback for speech training has also been researched in the context of clinical populations, notably in children with residual speech errors (i.e., errors that persist beyond the typical developmental stage) affecting English /ɹ/. Positive outcomes in this population have been reported in connection with both visual-acoustic biofeedback (e.g., McAllister Byun, 2017; McAllister Byun and Campbell, 2016; McAllister Byun and Hitchcock, 2012) and ultrasound biofeedback (e.g., Bacsfalvi and Bernhardt, 2011; Bernhardt *et al.*, 2005; McAllister Byun *et al.*, 2014;

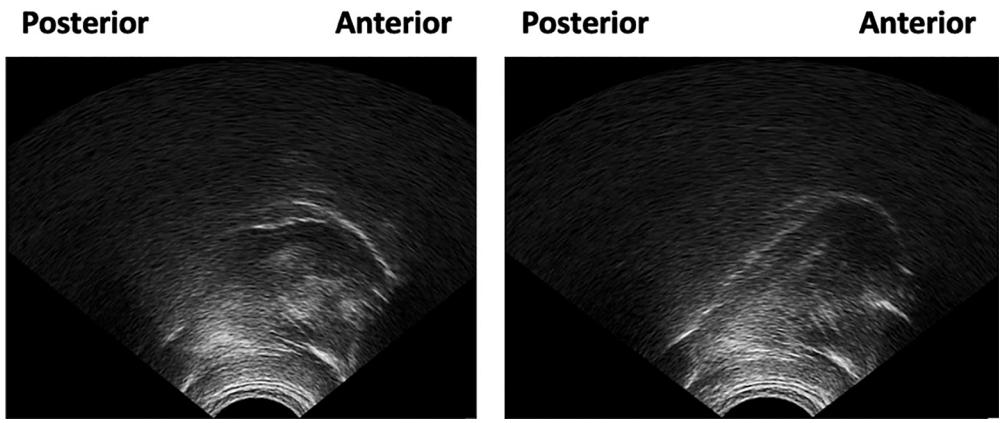

FIG. 2. Examples of ultrasound biofeedback displays in which the white line represents the surface of the tongue. Incorrect (left) and correct (right) productions of /y/ differ in the shape and position of the anterior and posterior tongue.

Preston *et al.*, 2016; Preston *et al.*, 2018, 2019). Despite finding generally beneficial effects of biofeedback training, these clinical studies have also described heterogeneity of response across individuals, with most studies reporting a mix of strong responders and non-responders (e.g., McAllister Byun *et al.*, 2014; Preston *et al.*, 2014). For instance, in their study of ultrasound visual biofeedback for residual speech errors, Preston *et al.* (2014) found that participants differed widely in both the degree of progress attained and the stage of training during which improvements became apparent. Likewise, in their study of ultrasound biofeedback intervention for /r/ misarticulation, McAllister Byun *et al.* (2014) observed strong gains in roughly half of the participants and minimal change in the other half. A high degree of variation between individuals was also reported in Kartushina and colleagues' (2015) study of visual-acoustic biofeedback training for L2 vowel production.

Authors of the above studies have posited that participant-level differences in motor skill, sensory acuity, or other factors might mediate the observed heterogeneity in outcomes. This possibility is aligned with research in the personalized learning framework (Perrachione *et al.*, 2011; Wong *et al.*, 2017), a line of investigation primarily associated with the L2 learning literature. Personalized learning suggests that speech outcomes can be optimized by measuring individual differences in learning performance, identifying factors that are predictive of such differences, and using these predictors to assign learners to a training paradigm that is optimally aligned with their ability profile (Wong *et al.*, 2017). An example of personalized learning in speech comes from Wong and Perrachione (2007), who found that successful and unsuccessful participants in an L2 tone learning task differed in their pre-training ability to identify pitch patterns in a non-linguistic context. Another study (Perrachione *et al.*, 2011) explored the relationship between pitch perception acuity and learning outcomes in a non-native lexical tone identification task in two training paradigms: a high-variability condition with multiple talkers producing the phonological contrasts to be learned, and a low-variability condition with only one talker. Perrachione *et al.* (2011) found that participants with high pre-training pitch perception acuity responded better to high-variability than low-variability training, while those with lower acuity showed the reverse pattern. This finding highlights the importance of considering individual differences when assigning learners to training paradigms. The present study aimed to identify characteristics that could be predictive of response or non-response to L2 production training incorporating different types of visual biofeedback.

B. Predictors of L2 production ability

Previous research has identified a number of factors that may play a role in predicting learners' success in producing the sounds of an L2. A range of studies have suggested that individuals with higher musicality or musical training may be more successful in L2 pronunciation (Christiner and Reiterer, 2015; Hu *et al.*, 2013; Milovanov *et al.*, 2010; Slevc and Miyake, 2006). Multiple studies have also suggested that greater success in L2 production may be associated with greater working memory capacity (Reiterer *et al.*, 2011; Rota and Reiterer, 2009) and/or phonetic or phonological awareness (Hu *et al.*, 2013; Nushi Kochaksaraie and Makiabadi, 2018). Other factors, such as motivation (Dörnyei, 1998) and empathy (Hu *et al.*, 2013), have also been implicated. In the neuroscience literature, which tends to emphasize the sensorimotor aspects of speech production, there has been particular focus on sensory processing as a predictor of L2 production capability. For instance, Simmonds *et al.* (2011) found greater activation in auditory and somatosensory brain regions during L2 production relative to L1 production and/or rest, while Reiterer *et al.* (2011) found greater activation in premotor and sensory regions in low-skill versus high-skill imitators of an unfamiliar language. Building on this line of research, the present study examined whether behavioral measures of auditory and somatosensory acuity would predict success in an L2 training task.

Theoretical models of L2 learning generally assume a close relationship between speech perception and production (Flege, 1995). Thus, differences in auditory acuity would seem to be a plausible candidate to explain individual differences in response to biofeedback training for L2 speech production. On the other hand, previous empirical research examining relationships between perceptual acuity and production accuracy in L2 learning has yielded mixed results (Bradlow *et al.*, 1997; Flege, 1995; Hanulíková *et al.*, 2012; Hattori and Iverson, 2010; Kartushina *et al.*, 2015; Nagle, 2018; Okuno and Hardison, 2016; Peperkamp and Bouchon, 2011). Some studies report no link between perception and production in L2 (Kartushina and Frauenfelder, 2014; Peperkamp and Bouchon, 2011), while other studies report weak to moderate correlations with large variation between individuals (Bradlow *et al.*, 1997; Nagle, 2018; Hanulíková *et al.*, 2012). For example, in a widely cited study by Bradlow *et al.* (1997), native Japanese learners of English were trained on perception of the English /ɹ/-/l/ contrast; no production training was provided. At the group level, there was evidence of improvement in production, as well as perception, supporting a link between the two domains. However, scores reflecting the magnitude of change from pre- to post-training in perception and production were not correlated across individual participants. These mixed results suggest that it is overly simplistic to try to reduce perception-production relations to a single direct link (Nagle, 2018).

In their study examining perception and production ability in L1 Spanish speakers learning French as an L2, Kartushina and Frauenfelder (2014) found that the best predictor of production accuracy was not an explicit measure of perceptual acuity but rather the compactness or stability of a given speaker's acoustic output across multiple repetitions of an L1 vowel. Specifically, they found that speakers who exhibited lower variability across repeated productions of an L1 vowel also produced L2 vowels more accurately. A similar result was reported by Kartushina *et al.* (2016) in a study of French speakers' acquisition of L2 vowels. One possible interpretation of these findings holds that individuals with less variable

production should also have more "blank space" between vowel categories (Kartushina and Frauenfelder, 2014). They may therefore be better able to form a distinct representation for producing and perceiving a novel L2 vowel category, rather than assimilating it to a native vowel. While variability is not a direct measure of perceptual acuity, studies such as Perkell *et al*. (2004a) and Franken *et al*. (2017) have found significant correlations between production variability and perceptual discrimination acuity for L1 vowels. However, variability in production is also affected by other factors, most notably articulatory stability (Kartushina and Frauenfelder, 2014). In sum, production variability is a factor related but not identical to perceptual acuity that may be particularly valuable in predicting individual differences in L2 phonetic learning.

The notion of links between variability in production and acuity in perception evokes research in the DIVA theoretical framework (Guenther, 1994, 1995, 2016; Guenther *et al*., 2006; Guenther *et al*., 1998).[1] Positing that the targets of speech are time-varying multidimensional regions in sensory space, the DIVA model predicts that speakers who represent a given speech sound with a narrower region in this sensory space should also be more precise in their phonetic realization of the sound in question. Empirical research in recent decades has accumulated a body of evidence that individual variation in speech production does indeed correlate with individual differences in auditory acuity (Newman, 2003; Perkell *et al*., 2004a; Villacorta *et al*., 2007). However, it is important to consider that in DIVA and related models, the targets of speech are somatosensory, as well as auditory, in nature. Empirical research supports the importance of the somatosensory domain in speech learning (Borrie and Schäfer, 2015), including multiple studies that have replicated a finding of group differences in oral somatosensory acuity between adolescents with residual speech errors and age-matched typical speakers (Fucci, 1972; Fucci and Robertson, 1971; McNutt, 1977). Acuity in auditory and somatosensory domains appear to be independent of one another (Fucci, 1972; Nasir and Ostry, 2008) and, in fact, it has been specifically argued that speakers may show a "sensory preference" to compensate for auditory versus somatosensory perturbations in speech adaptation tasks (Lametti *et al*., 2012). Therefore, one possible explanation for the mixed results in previous studies of perception-production links in L2 learning is that these studies evaluated sensory acuity only in the auditory domain. Cases of dissociation could, in principle, be a reflection of relative strength or weakness in the unmeasured somatosensory domain.

These theoretical and empirical considerations, along with neuroimaging findings (e.g., Reiterer *et al*., 2011), suggest that even among typical adult speakers, differences in auditory and somatosensory acuity may influence performance in speech learning tasks. If a learner has relatively low sensitivity in the auditory domain, they may have difficulty establishing a distinct target for a novel sound in auditory-perceptual space. An individual with relatively low somatosensory ability, meanwhile, would be expected to have difficulty establishing precise somatosensory targets for accurate production, which correspond to tactile-kinesthetic correlates of articulator placement. Thus, even if they hear the target sound accurately, they may have difficulty determining what movements of the articulators are needed to achieve a particular auditory target. In keeping with the personalized learning framework, such differences could have implications for the selection of a training paradigm in the context of L2 learning. Specifically, we hypothesize that speakers will derive the greatest benefit from training if they receive enhanced feedback in a domain where their intrinsic sensory acuity is weaker, and they will show less effect of training if the enhanced feedback is directed at a domain where their acuity is already strong. In visual-acoustic biofeedback, the real-time signal tracks directly with the auditory-acoustic aspects of speech and is only indirectly related to articulator movements. By contrast, ultrasound biofeedback provides a direct display of articulator placement that bears only an indirect relation to auditory-acoustic outcomes. Thus, we hypothesize that learners with low auditory acuity should derive the greatest benefit from visual-acoustic biofeedback, which provides a clearly defined visual display indicating how close the speaker's acoustic output is to the target sound. Meanwhile, a learner with relatively poor somatosensory sensitivity can be hypothesized to benefit most from the visual information about articulator placement provided by ultrasound biofeedback.

We investigated this possibility using a task in which native English speakers received visual-acoustic or ultrasound biofeedback training targeting phonetically accurate production of two Mandarin vowels, /u/ and /y/. Participants were randomly assigned to a single biofeedback condition that was then used to train both vowels. Based on the reasoning laid out above, we hypothesized that individual differences in auditory and somatosensory acuity would interact with biofeedback modality to predict learning outcomes in this task. Specifically, we hypothesized that the magnitude of speakers' response to visual-acoustic biofeedback would be significantly negatively correlated with a measure of auditory acuity (i.e., those speakers with low auditory acuity would derive greater benefit from visual-acoustic biofeedback training), while the magnitude of response to ultrasound biofeedback would correlate negatively with a measure of oral somatosensory acuity. We did not expect a significant correlation between acuity in a domain and response to biofeedback that targets a different domain (e.g., auditory acuity and ultrasound biofeedback), leading us to predict a significant interaction between sensory acuity and biofeedback type.

Although our primary hypothesis pertains to sensory acuity as a predictor of response to visual-acoustic and ultrasound biofeedback, we measured other properties that could also play a significant role. In addition to their findings about pitch perception, Perrachione *et al*. (2011) reported that individual differences in phonological awareness could explain significant variance in adult learners' success in acquiring non-native tones. This led us to measure phonological awareness as a potential predictor of outcomes in our L2 learning task. We also examined the effect of accuracy in vowel imitation at baseline, since previous literature has reported that individuals who produce more accurate imitations of non-native speech sounds prior to training typically show a smaller magnitude of improvement over the course

of training than those who began with less accurate productions (Bradlow *et al.*, 1997; Kartushina *et al.*, 2015). Finally, like Kartushina and Frauenfelder (2014), we measured token-to-token variability in baseline production as a potential predictor of response to training—although we remain agnostic as to whether this measure reflects auditory acuity, articulatory stability, or some combination thereof. Unlike Kartushina and Frauenfelder (2014), we measured variability in L2 imitation rather than L1 production, reasoning that variability should be correlated across languages.[2] Kartushina and Frauenfelder (2014) did not find a significant correlation between compactness in L1 and L2, but they did note that they used different tasks to measure compactness in the two languages and acknowledged that such a correlation would be theoretically plausible. Moreover, Kartushina *et al.* (2015) found a significant correlation between variability and accuracy in L2, such that speakers with low variability in L2 production also produced closer approximations of L2 targets both before and after training.

A final factor to consider when measuring L2 learning outcomes is the identity of the speech sound(s) being targeted. Participants in the present study were trained to produce two Mandarin vowels, /u/ and /y/. Phonetically, both /y/ and /u/ are characterized by a low first formant frequency (F1), associated with high tongue placement; they contrast in the frequency of the second formant (F2), which is high in the front vowel /y/ but low in the back vowel /u/. Both vowels are produced with rounding of the lips. Previous literature has hypothesized that speakers process L2 sounds differently based on perceptual similarity to established L1 categories. L2 sounds that are perceived as similar to others in the L1 inventory may be assimilated to the closest native category, while L2 sounds that have no close match in the L1 inventory may remain "uncategorized" (e.g., Best, 1995; Flege, 1995). Relative to the English vowel inventory, the high front rounded vowel /y/ does not have a close counterpart and is thus likely to be treated perceptually as uncategorized (Chang *et al.*, 2011; Levy and Strange, 2008; but cf. Best *et al.*, 2003). Mandarin /u/, while phonetically more back than English /u/ (Chen *et al.*, 2001), is perceptually similar to English /u/ and thus expected to behave as an "assimilated" vowel in perception and production. The finding by Kartushina and Frauenfelder (2014) that an assimilated vowel contrast was both perceived and produced more accurately than an uncategorized contrast suggests that Mandarin /u/ might be produced with greater accuracy than /y/. On the other hand, as noted above, the magnitude of improvement over the course of training may be larger when productions at baseline are less accurate (Bradlow *et al.*, 1997; Kartushina *et al.*, 2015). Thus, the identity of the trained vowel could play a role in determining both baseline accuracy and learning outcomes.

In summary, the present study aimed to answer the following questions: (1) Do English-speaking adults significantly improve their production of the Mandarin vowels /y/ and /u/ over a brief period (30 min per vowel) of training enhanced with either visual-acoustic or ultrasound biofeedback? (2) Do individual profiles of sensory acuity interact with biofeedback type in determining response to training? We predicted that learners would, on average, show progress in production accuracy over the course of training. We further hypothesized that individuals with relatively low auditory acuity would show a greater magnitude of response to visual-acoustic than ultrasound biofeedback, while individuals with relatively low somatosensory acuity would show the reverse pattern.

II. METHODS

A. Participants

This project was approved by the Institutional Review Board at New York University (protocol IRB-FY2016-583); all participants signed an informed consent form before starting the study. Sixty-five female English speakers aged between 18 and 30 years old participated in the study.[3] All participants completed an online questionnaire prior to the experiment to report their language background and speech-language history. Participants were included only if they reported no history of speech and language disorders. They were also required to report healthy dental status (e.g., no missing teeth), since it has been suggested that dental status could affect oral stereognosis (Jacobs *et al.*, 1998). They were also asked to report all languages to which they were exposed. All participants were required to be native speakers of American English. Participants could be bilingual or multilingual, but individuals who reported exposure to Mandarin or another language that features a front-back contrast in high rounded vowels (e.g., French, German, or Swedish) were excluded. Participants were also required to self-report normal hearing ability, and they additionally completed a pure-tone hearing screening (20 dB at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz) to confirm normal hearing status. All participants were asked to complete two sessions on two separate days and were compensated $20 per session. An additional four participants were recruited but did not complete the study for reasons including malfunction of the experimental equipment ($n = 1$), reporting previous knowledge of French that was not identified at pre-screening ($n = 1$), and failure to attend the second session ($n = 2$). Data from participants who did not complete both sessions were not included in the analyses reported here. Table I provides a summary of self-reported linguistic influences (languages spoken; geographic region judged to have the greatest impact on speech patterns) for all included participants.[4] The questionnaire used to evaluate participants' speech and language history is available at https://osf.io/b6qux/.

B. Session 1: Assessment

Session 1 began with the hearing screening mentioned previously. Participants then completed tasks assessing auditory and oral somatosensory acuity as well as phonological awareness. We describe each measure in greater detail below.

1. Auditory acuity

Auditory acuity in the specific context of the Mandarin /y/-/u/ contrast was assessed with an AXB discrimination task. A synthesized continuum from /y/ to /u/ was generated

TABLE I. Summary of linguistic influences reported by included participants. Geographic region of origin was elicited with the question "Which state or region do you think has had the biggest impact on your accent?".

	Number of participants	Detail
Languages spoken		
Speak another language with some fluency	18	Spanish ($n = 11$), Korean ($n = 5$), Tagalog ($n = 1$), Vietnamese ($n = 1$)
Some exposure to another language	41	Spanish ($n = 38$), Italian ($n = 3$), Hebrew ($n = 3$), Latin ($n = 3$), Japanese ($n = 2$), Korean ($n = 1$)
Speak only English	6	
Geographic region of origin		
Northeast	30	
Pacific West	12	
Midwest	9	
Southwest	4	
Southeast	9	
No response	1	

from a native speaker's productions of isolated /y/ and /u/ vowels using the speech algorithm STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram; Kawahara *et al.*, 2013). The algorithm created 240 equally spaced steps between the endpoints of the continuum, adjusting the heights of all formants. In each trial of the AXB task, participants listened to three stimuli over headphones (Sennheiser HD 429, Sennheiser, Wedemark, Germany) and were asked to report which of the first (A) or last (B) stimuli was identical to the center stimulus (X). A gamified interface, in which participants received points for correct responses, was used to support sustained attention over this extended task. Following Villacorta *et al.* (2007), an adaptive staircase procedure was used to determine the just noticeable distance (JND) score for the detection of a difference in vowel formants. The procedure used a one-up, two-down protocol to adjust the stimulus interval following the response to each trial, initially decreasing the stimulus distance by eight steps after each correct response and increasing the stimulus distance by four steps after each incorrect response. Following every fourth reversal in the direction of stimulus change, the size of the stimulus adjustment was reduced by half in order to gain an increasingly precise estimate of the JND. The task ended after 12 reversals or after a maximum of 80 trials. The mean distance between the stimuli at the final four reversals was used as an estimate of participants' JND (henceforth, auditory JND), where a smaller JND corresponds with a higher degree of acuity in auditory perception. Participants completed the AXB task twice, but due to evidence of task learning effects (i.e., a substantial proportion of participants who achieved ceiling-level performance in the second run[5]), only the results from the first run were used in the analyses reported here.

2. Somatosensory acuity

Oral somatosensory acuity was assessed using a spatial resolution acuity task (Steele *et al.*, 2014), which required participants to use their tongue to identify letters embossed on Teflon (Chemours, Wilmington, DE) strips. Although it is a non-speech measure, its potential relevance to speech is established by previous research reporting group differences between individuals with and without residual speech errors in similar oral stereognosis tasks (Fucci, 1972; Fucci and Robertson, 1971). Materials, which were manufactured per specifications identical to those used in Steele *et al.* (2014), were Teflon strips embossed with capital letters (A, I, J, L, T, U, and W) of seven different sizes (2.5, 3, 4, 5, 6, 7, and 8 mm). Participants were required to sit in front of the investigator, wearing sunglasses to avoid seeing the letters. The investigator delivered each strip to the participant, who then placed the strip in the mouth with the embossed side facing down. The participant was instructed to identify the letter by searching with the tongue tip, and verbally let the investigator know his/her answer. Stimuli were presented in a staircase fashion, starting with a strip with a randomly chosen letter of medium size (i.e., 5 mm). A correct identification led to a one-step decrease in size and an incorrect identification led to a one-step increase in size on the next attempt. The task ended after 8 reversals or a maximum of 28 trials. Following Steele *et al.* (2014), the outcome measure was the mean letter size (henceforth, somatosensory MLS) across all correctly answered trials; a smaller MLS is suggestive of a higher degree of acuity in somatosensory perception. Participants completed the task twice, but as in the auditory ABX task, data from the second run were discarded due to concerns of a learning effect, since many participants achieved a perfect score on their second attempt at the task. Participants' apparent learning of the task was attributed to the small number of letters included in the stimulus set; we return to this task limitation in Sec. IV (Discussion).

3. Phonological awareness

Phonological awareness was assessed using the Phonological Awareness Composite Score (PACS) derived from performance across three subtests of the *Comprehensive Test of Phonological Processing—Second Edition (CTOPP-2)*: Elision, Blending and Phoneme Isolation (Wagner *et al.*, 2013). The elision task requires participants to take out segments from spoken words to form new words, while the blending task instructs them to combine individual sounds together to form words, and the phoneme isolation task requires them to identify a sound at a specific position within a word.

C. Session 2: Vowel production training

1. Stimuli

As described above, our primary task was imitative production of the Mandarin vowels /u/ and /y/, elicited both before and after training. The audio stimuli used in imitative elicitation and biofeedback training were productions of /y/ and /u/ in isolation, recorded by three female native speakers of Mandarin who produced six tokens of each vowel.[6]

Recordings took place in a sound-attenuated booth on a Computerized Speech Lab (CSL) system (model 4500, PENTAX Medical, Montvale, NJ) with a 44 kHz sampling rate and 16-bit encoding. Speakers produced the target vowels into a table-mounted Shure SM48 microphone (Shure, Niles, IL) with a five-inch mouth-to-microphone distance, and recordings were saved as uncompressed WAV files.

2. Training

Participants received 30 min of production training for each of the two vowels, /u/ and /y/. The order in which vowels were trained was counterbalanced across participants. As noted previously, participants were randomly assigned to receive either visual-acoustic biofeedback or ultrasound biofeedback with the same biofeedback modality for both vowels. We briefly summarize the methods used below, but the complete protocol used in each condition can be found at https://osf.io/b6qux/. The production training approach adopted in this study was derived from previous work using biofeedback to elicit target sounds in children with speech sound disorder (e.g., for visual-acoustic biofeedback, McAllister Byun, 2017; for ultrasound biofeedback, Preston *et al.*, 2019).

Prior to biofeedback training, each participant needed to be matched to a native speaker whose template (either a formant template or a tongue shape template, depending on the biofeedback modality) would be used as a target during production training. This was accomplished by judging which of the three native speakers of Mandarin was most similar to the participant in formant frequencies for a non-target sound. For this purpose, we used the PENTAX Medical Sona-Match real-time Linear Predictive Coding (LPC) spectrum module on the CSL system. (Note that the visual-acoustic biofeedback display was used in the process of selecting a target speaker even for participants assigned to the ultrasound biofeedback condition.) Participants were instructed to sustain /i/ vowels while templates from different control speakers were superimposed, and the best match was selected as that participant's target speaker. The vowel /i/ was used because it has comparable phonetic properties across English and Mandarin, and it was not a target vowel in the present study. During template selection (and practice, for individuals assigned to the visual-acoustic biofeedback condition), we used the Sona-Match standard "real-time LPC window for females" setting (11 025 Hz sampling rate, 12-order LPC in a 3500 Hz view range) for all participants.

After a target speaker was identified, the investigator familiarized the participant with the biofeedback modality to be used in training. For participants in the visual-acoustic biofeedback condition, the above-described process of matching an /i/ template to identify a target talker doubled as an initial familiarization with the biofeedback device, in which participants learned that different speech sounds have different formant peaks and changes in vocal tract configuration can cause these peaks to move around. Following this training, all participants were judged to demonstrate adequate understanding of the relationship between vocal tract manipulations and changes in formant frequency. In the ultrasound training condition, a separate brief training was provided to familiarize the participant with basics such as the orientation of the tongue on the screen and the correct way to hold the ultrasound probe.[7] The ultrasound imaging system used was a Siemens Acuson X300 (Siemens, Munich, Germany) with a C8-5 wideband curved array transducer (frequency range 3.1–8.8 MHz, 25.6 mm footprint, 109 deg field of view; scanning settings were not adjusted on a per-participant basis for the present study). The investigator also demonstrated how to use the ultrasound template, a transparency with a trace of the target speaker's tongue shape superimposed on the ultrasound screen. A tongue shape template for /i/ was used for training purposes.

After this general introduction, participants began training for the first target vowel, randomly assigned to be /u/ or /y/. Training began with a few minutes of unstructured practice in which the participant was encouraged to repeat the target sound while attempting to make the real-time visual display of their own production align with the target speaker's template, which represented formant peaks in the visual-acoustic biofeedback condition and tongue shape in the ultrasound biofeedback condition. During this process, the first author (J.J.L.), a native speaker of Mandarin, provided cues and general encouragement. Specific articulator placement cues (e.g., "try moving your tongue forward more") were used in both the visual-acoustic and the ultrasound biofeedback conditions. This was intended to keep the verbal instructions as consistent as possible across conditions, such that the only difference between conditions was the type of visual biofeedback provided. Participants then completed 30 min of structured training for the selected vowel. During this training, participants were instructed to repeat audio models presented through headphones (isolated vowel tokens produced by their target talker) while simultaneously trying to match their visual feedback display with the corresponding visual-acoustic or ultrasound template. The training consisted of 20 blocks of 6 trials for a total of 120 trials. Qualitative feedback, including cues for articulator placement as described above, was provided after each block. Participants then completed another 30 min of training for the second vowel, following the same procedures described for the first training session.

3. Probes

Participants' production of the target vowels was assessed with an imitative probe task. Audio models from the matched control speaker, randomized at the token level, were presented through headphones (Sennheiser HD 429, Sennheiser, Wedemark, Germany), and the participant repeated after each token, advancing in an administrator-paced fashion. Twenty repetitions of each vowel, elicited in a blocked fashion, were audio-recorded with the same parameters described above for control speakers. A midpoint probe measure was collected after training for the first vowel, and a post-training probe was collected after training for the second vowel. Both midpoint probes and post-training probes were identical to the pre-training probe, eliciting 20 productions of each vowel in a blocked fashion. In midpoint and post-training probes, the vowel that was most recently trained was elicited first.

D. Measurement

Formant measurements from the midpoint of each vowel produced by the participants in pre-, mid-, and post-training probes were extracted using an automated process in Praat (Boersma and Weenink, 2016). The onset and offset of each isolated vowel production were automatically detected using a -25 dB threshold to differentiate silent versus sounding intervals. A Praat script (Lennes, 2003) was then used to extract measurements of the first three formants from a 0.05s window around the selected point in each word, although only F1 and F2 were used in the analyses that follow. Individualized LPC filter order settings were selected for each speaker, using visual inspection to optimize agreement between automated formant tracks and visible areas of energy concentration on the spectrogram. J.J.L. and S.A. independently selected an LPC order for each participant and then compared their selections; cases of disagreement were resolved by consensus. This resulted in a total of 39 individuals whose data were analyzed with an order 11 filter, 24 with an order 10 filter, and 2 with an order 9 filter. F1 and F2 values from the target tokens produced by native Mandarin speakers were obtained using the same protocol. After extraction, formant frequencies were transformed into the psychoacoustic Bark scale using the vowels package in *R* (Kendall *et al.*, 2018).

Euclidean distance (ED) from the center of the distribution of productions by the participant's target talker was used as our primary measure of production accuracy.[8] A smaller ED value indicates that the participant's production is closer to the target distribution, suggesting higher accuracy. ED was calculated for each token, and median ED across 20 repetitions was calculated for each participant and vowel at baseline and immediate post-training time points.[9] Change in ED from baseline to post-training was also calculated by subtracting baseline median ED from post-training median ED. A negative value for change in median ED was considered indicative of improvement over the course of training. Figure 3 shows the distribution of productions of /y/ for one sample participant [identification (ID) number 1026] at pre- and post-training time points, as well as the center of the distribution of tokens produced by the native Mandarin speaker who served as the target talker for this participant. This participant's productions were closer to the target distribution after training, reflected in a decrease in mean ED. This change over the course of training is also perceptually apparent in Mm. (1), which includes the target talker's production of /y/ and imitative productions by the participant both before and after training.

Mm. 1. Productions of /y/ by (a) Mandarin native speaker, (b) sample participant 1026 before training, and (c) sample participant 1026 after training. This is a file of type "wav" (299 KB).

We also calculated the area of an ellipse representing the 95% confidence interval around the multivariate mean of the distribution of a subject's productions for each vowel and time point. This provides an index of token-to-token variability in a speaker's productions and was thus treated as analogous to the compactness measure used by Kartushina and Frauenfelder (2014).[10] The distributions in Fig. 3 also contrast in the area of the ellipse, with a smaller area reflecting lower variability in production after training.

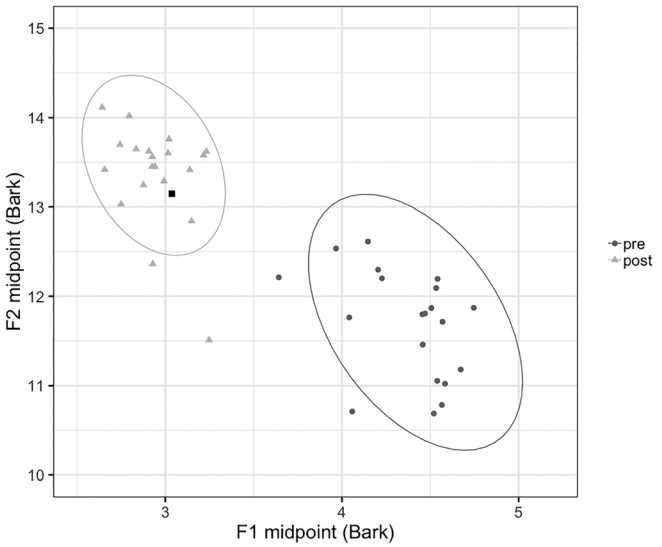

FIG. 3. Distribution of productions of /y/ from one sample participant (1026) with Bark-transformed F1 on the *x* axis and Bark-transformed F2 on the *y* axis. Circles represent productions before training targeting /y/, and triangles represent productions after training. The mean value for the L1 target speaker is displayed for comparison (black square).

Here, and throughout the paper, all computations and data visualization were carried out in the *R* software environment (R Core Team, 2015). Data wrangling and plotting were completed with the packages tidyr (Wickham, 2016), dplyr (Wickham and Francois, 2015), and ggplot2 (Wickham, 2009). Complete data and code to reproduce all figures and analyses in the paper can be found at https://osf.io/9djpm/.

E. Data cleaning

After ED was calculated, the dataset was cleaned for outliers. Outlier data points were identified as ED values that fell at least two median absolute deviations above or below the median ED for a given individual speaker, vowel, and time point (pre or post).[11] A total of 231 data points (4.5% of the total) were eliminated in this way. In addition, individual participants' results were compared against measures of central tendency from the complete group of subjects. A total of 9 subjects were found to have a median ED that fell more than 2 standard deviations (SDs) away from the group mean ED,[12] and a total of 12 fell more than 2 SD away from the group mean area of the ellipse. To be conservative, only the five individuals who were outliers with regard to both mean ED and variability (i.e., area of the ellipse) were eliminated from the data set. The included set of 60 individuals had a mean age of 20.03 years old (SD = 1.93 yr; range = 18–26 yr).

III. RESULTS

A. Descriptive statistics and preliminary analyses: Baseline measures

In this section, we report summary statistics for each of the measures obtained at baseline and also test for correlations

between these variables that could affect analyses by introducing multicollinearity. The baseline measures investigated in this section include auditory JND, somatosensory MLS, PACS, and variability (area of the ellipse) for each vowel.

Descriptive statistics for all of the above-listed baseline measures are reported in Table II. The ranges and SDs suggest that scores for each measure were reasonably dispersed across participants. However, the results of Shapiro-Wilk tests for normality reported in Table II indicate that several of the variables were not normally distributed. Non-normality was particularly pronounced in the case of auditory JND, which showed some sign of a ceiling effect. Accordingly, methods requiring strict parametric assumptions will be avoided for analyses involving these measures.

Spearman's rho was used to test for problematic correlations among measures of auditory JND, somatosensory MLS, and PACS. All correlations were nonsignificant [JND and MLS, $\rho(58) = 0.05$, $p = 0.7$; JND and PACS, $\rho(58) = -0.17$, $p = 0.2$; MLS and PACS, $\rho(58) = 0.02$, $p = 0.9$]. Spearman's rho was also used to test for correlations between the auditory, somatosensory, and PACS measures with variability (area of the ellipse) at baseline. There were no significant correlations between variability and auditory JND [$\rho(58) = 0.01$, $p = 0.95$], somatosensory MLS [$\rho(59) = -0.01$, $p = 0.96$], or PACS [$\rho(59) = 0.12$, $p = 0.37$]. These findings did not change when data from the target vowels /y/ and /u/ were analyzed separately. Finally, a series of Wilcoxon tests was carried out to examine whether the various baseline predictors differed between the groups of participants randomly assigned to receive visual-acoustic versus ultrasound biofeedback training. The groups were not found to differ significantly with respect to auditory JND ($W = 476.5$, $p = 0.69$), somatosensory MLS ($W = 418$, $p = 0.65$), PACS ($W = 484$, $p = 0.61$), or variability ($W = 361$, $p = 0.19$). In summary, the examination of baseline predictors revealed no problematic correlations between measures or discrepancies between the groups randomly assigned to different biofeedback conditions.

B. Descriptive statistics and preliminary analyses: Change in ED and variability

This section reports summary statistics for measures of change in ED and variability over the course of training. Group comparisons are used to assess whether the magnitude of change differed over various subdivisions of the data, including target vowel, type of biofeedback, and the order in which target vowels were trained.

TABLE II. Descriptive statistics for baseline measures.

Measure	Median (MAD)	Range	Shapiro-Wilk p-value
Auditory JND (continuum steps)	19.5 (19.6)	2–58	<0.001
Somatosensory MLS (mm)	4.23 (1.1)	3–7.8	<0.01
PACS (standard score)	105 (10.4)	77–133	0.15
Variability (area of the ellipse, Bark squared)	1.74 (0.9)	0.29–6.5	<0.001

The boxplots in Fig. 4 show the pre- and post-training distribution of ED values observed for each of the two target vowels, subdivided to reflect the type of biofeedback provided (upper panel) and the order in which vowels were targeted (lower panel).

Figure 4 suggests some degree of difference between the target vowels /y/ and /u/ with respect to baseline ED and change in ED over the course of training. An exploratory Wilcoxon test indicated that the difference between vowels was significant for baseline ED ($W = 2833$, $p < 0.001$). Contrary to expectations from previous literature (e.g., Kartushina and Frauenfelder, 2014), baseline accuracy was higher for the uncategorized vowel /y/ than the assimilated vowel /u/. The vowels did not differ significantly with respect to the magnitude of change in ED ($W = 1692$, $p = 0.57$). Both vowels showed a reduction in ED over the course of training; the effect size of the change was moderate in magnitude for both /u/ (Cohen's $d = -0.46$) and /y/ (Cohen's $d = -0.43$). Expressed as a percentage of baseline ED, the mean reduction in ED was 19.8% for the vowel /u/ and 23.6% for the vowel /y/. Neither biofeedback type nor vowel order appears to have had a marked impact on the general pattern observed in Fig. 4. A Wilcoxon test indicated that the groups randomly assigned to receive ultrasound versus visual-acoustic biofeedback did not differ with regard to ED at baseline ($W = 356$, $p = 0.17$). Examination of differences in the magnitude of change across biofeedback types is deferred until the regression models reported in Sec. III C. An additional Wilcoxon test confirmed that the group of participants who received training targeting /y/ first did not differ from the group who received training for /u/ first ($W = 452$, $p = 0.98$).

Figure 5 shows the equivalent data for our primary measure of variability in production, the area of the ellipse. Both vowels showed a reduction in variability over the course of the training; the effect size of the change (Cohen's d) was −0.21 for /u/ and −0.33 for /y/. Wilcoxon tests revealed that this difference in the magnitude of change between vowels was not statistically significant ($W = 1863$, $p = 0.74$), nor did the vowels differ significantly in baseline variability ($W = 1909$, $p = 0.57$). The groups randomly assigned to receive ultrasound versus visual-acoustic biofeedback did not differ with regard to mean variability at baseline ($W = 361$, $p = 0.19$). Finally, there was no difference in the change in variability associated with the order in which vowels were trained ($W = 539$, $p = 0.19$).

C. Regression models

Linear regression models were used for all inferential statistics. Linear regressions were favored over mixed effects models because, with the task and stimuli held constant from trial to trial, within-subject fluctuations in speech acoustics are better interpreted as measurement error than meaningful variation, and therefore each subject was represented by his/her median ED across trials. Due to the difference in ED between vowels at baseline and for ease of interpretation of regression results, data associated with the /u/ and /y/ vowels were analyzed separately. Two separate models were fit for each vowel: the first examined predictors of median ED at baseline, and the second examined predictors of change in

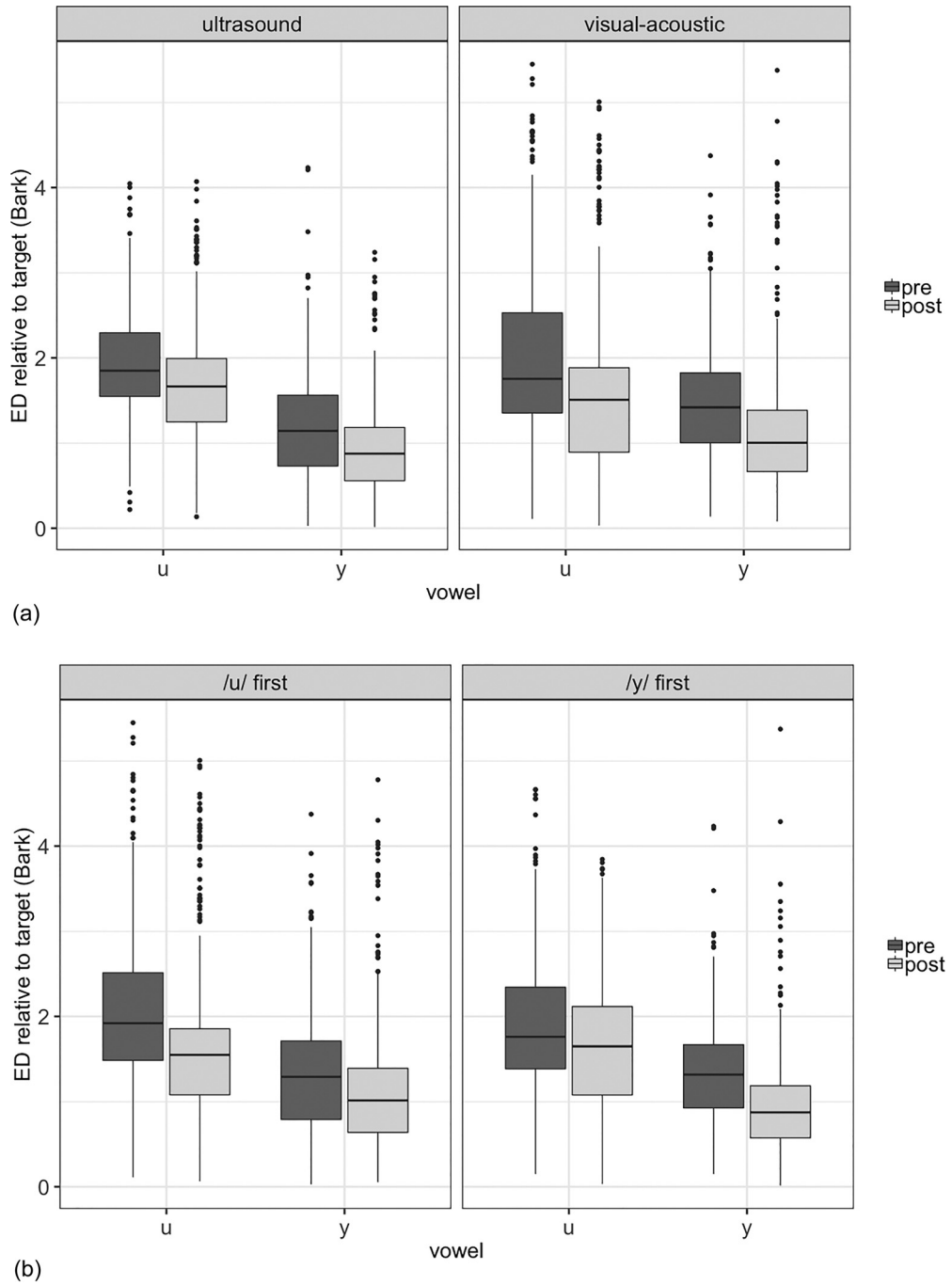

FIG. 4. ED before and after training, separated by vowel and biofeedback type (top) and by vowel and order of training (bottom).

median ED over the course of training. Both models included covariates of auditory JND, somatosensory MLS, PACS, and variability (area of the ellipse) at baseline. The model examining change in median ED additionally included covariates of variability at post-training, biofeedback condition (visual-acoustic versus ultrasound), and baseline accuracy (median ED relative to the native speaker distribution). Finally, the model also included the theoretically predicted interaction between biofeedback condition and sensory acuity (both auditory JND and somatosensory MLS).

In the first set of models, none of the independent variables examined were found to be significant predictors of participants' accuracy in imitating each of the non-native vowels /y/ and /u/ at baseline after controlling for all other variables in the model. This finding was contrary to our hypothesis that auditory JND would significantly predict participants' pre-training accuracy in imitating /y/ and /u/ vowels. In fact, the overall regression model was non-significant in each case [for /u/, $F(4,55) = 0.28$, $p = 0.89$; for /y/, $F(4,55) = 0.35$, $p = 0.84$], indicating that the models with all variables included did not account for significantly more variance than the intercept-only models. This suggests that variance in baseline accuracy might be influenced by additional factors that were not measured as part of the present study, a point we return to in Sec. IV (Discussion). Complete results of these regressions can be found at https://osf.io/jdz6s/.

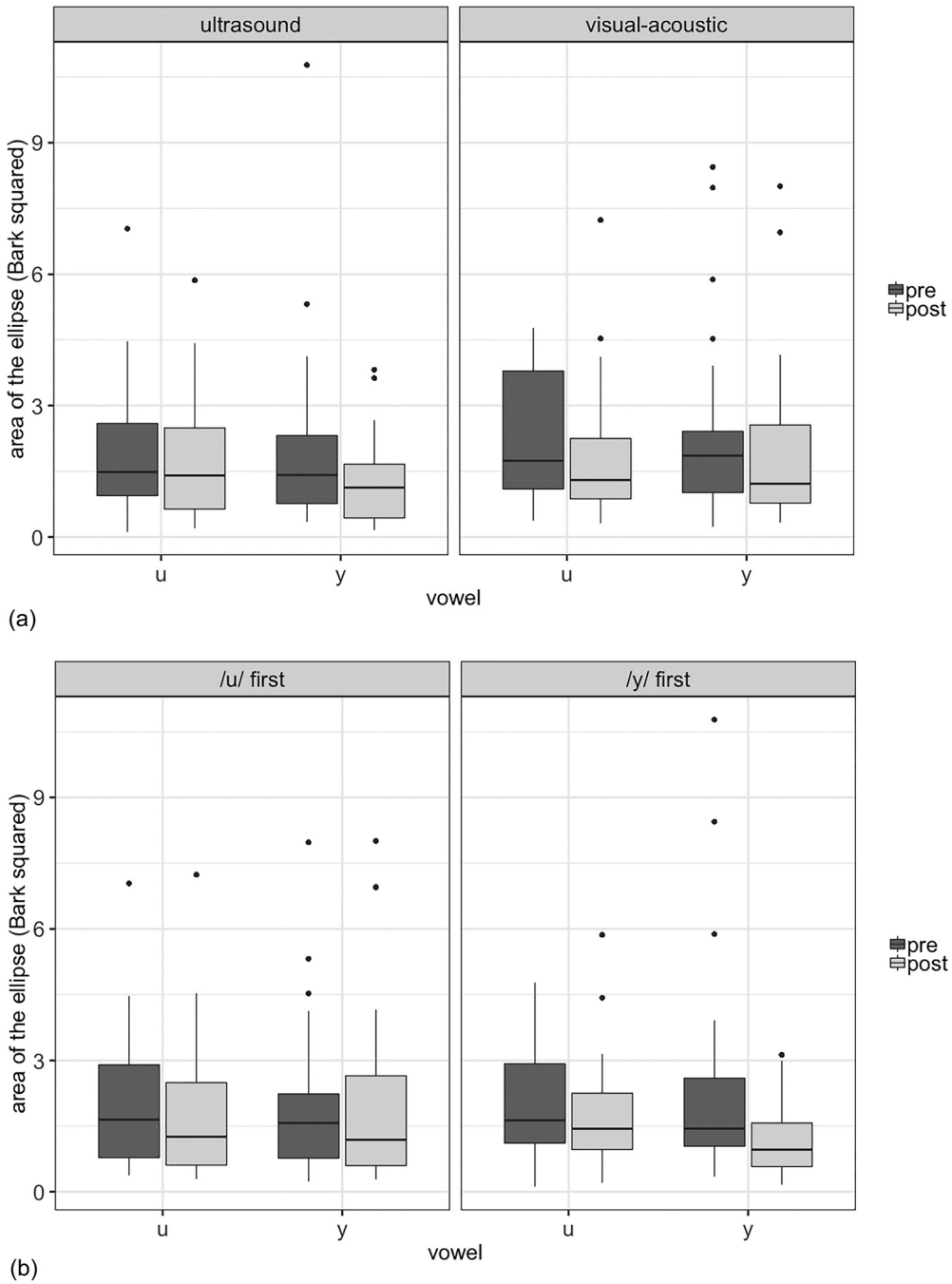

FIG. 5. Area of the ellipse before and after training, separated by vowel and biofeedback type (top) and by vowel and order of training (bottom).

The second set of models examined predictors of change in ED from pre- to post-training for the two vowels /u/ and /y/. Recall that an improvement in accuracy corresponds with a negative number in this context. Results of the model for /u/ yielded a significant main effect of median ED at baseline [$\beta = -0.47$, standard error (SE) $= 0.10$, $p < 0.001$], indicating that individuals who had a higher median ED prior to training tended to show a greater reduction in ED. (See Fig. 6.) There was also a significant effect of PACS score ($\beta = -0.02$, SE $= 0.0072$, $p = 0.01$), indicating that individuals with higher PACS scores tended to show a greater reduction in median ED (Fig. 7). For the vowel /y/, there was a significant effect of median ED at baseline ($\beta = -0.66$, SE $= 0.10$, $p < 0.001$), again with a negative coefficient (Fig. 6). In addition, for the /y/ vowel there was a significant main effect of variability (area of the ellipse) at the post-training time point ($\beta = 0.09$, SE $= 0.037$, $p = 0.02$). The effect of post-training variability has a positive coefficient, which suggests that a positive response to training (large reduction in median ED from pre- to post-training) tended to associate with a low degree of variability at the post-training time point (Fig. 8). Complete regression results for /y/ and /u/ are reported at https://osf.io/jdz6s/.

No significant effect of biofeedback type was found in either the model for /y/ ($\beta = 0.39$, SE $= 0.44$, $p = 0.38$) or /u/ ($\beta = -0.48$, SE $= 0.65$, $p = 0.46$). However, this failure to

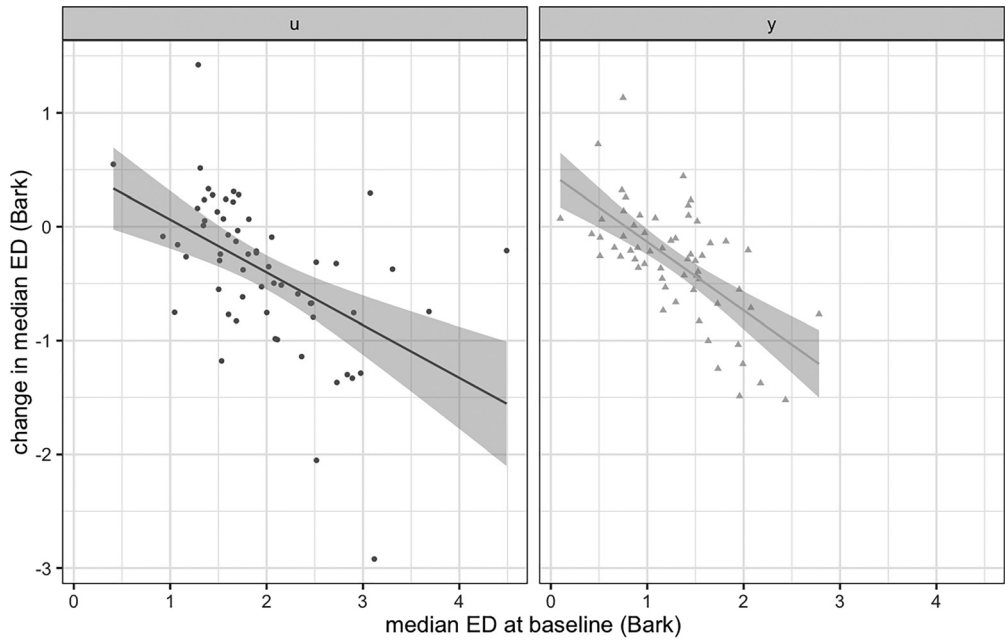

FIG. 6. Association between pre-training median ED and change in median ED over the course of training, partitioned by vowel.

reject the null hypothesis of no significant difference between biofeedback types does not, in itself, warrant the conclusion that both types are equally effective. We therefore conducted a *post hoc* equivalence test (Lakens, 2017), which treats non-equivalence between groups as the null hypothesis; vowels were pooled for the purpose of this analysis. The equivalence test was non-significant, $t(48.94) = -0.52$, $p = 0.30$, given equivalence bounds of -0.14 and 0.14 (on a raw scale) and an α of 0.05. Thus, although the regression did not show a significant difference between biofeedback types, the equivalence test did not support a conclusion that the magnitude of change was equivalent across the two conditions.

IV. DISCUSSION

A. Summary of hypotheses

The present study measured the magnitude of learning when native speakers of English were trained to produce two Mandarin vowels, /u/ and /y/, using either visual-acoustic or ultrasound biofeedback. Previous literature has indicated that individuals' responses to visual biofeedback training are highly heterogeneous, but the causes of this variability remain poorly understood. As suggested by the personalized learning framework (Wong *et al.*, 2017), identifying individual factors that are predictive of differences in learning outcomes could help maximize gains by pairing learners with a

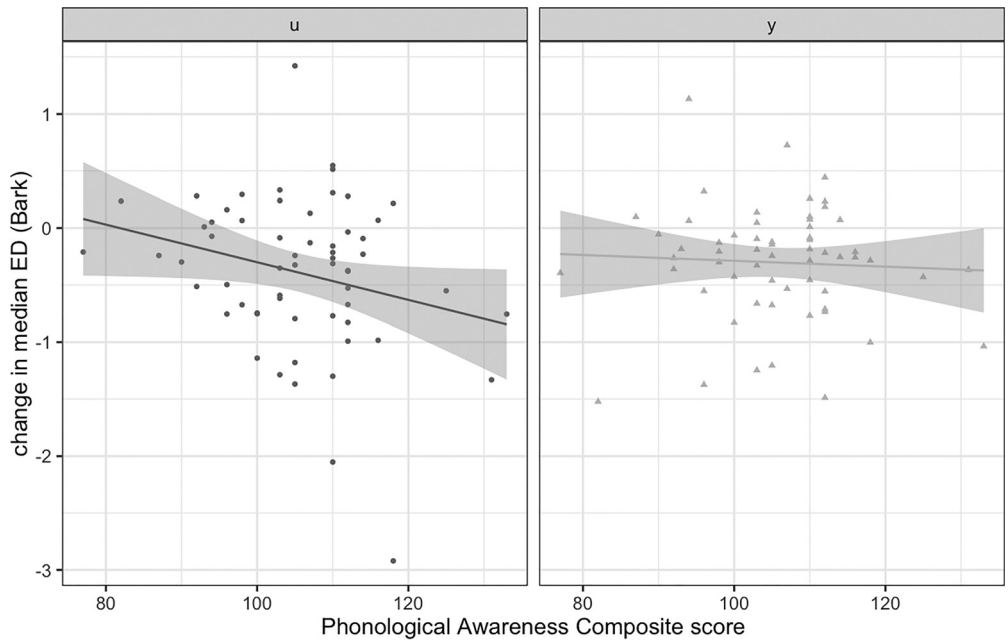

FIG. 7. Association between PACS score and change in median ED over the course of training, partitioned by vowel.

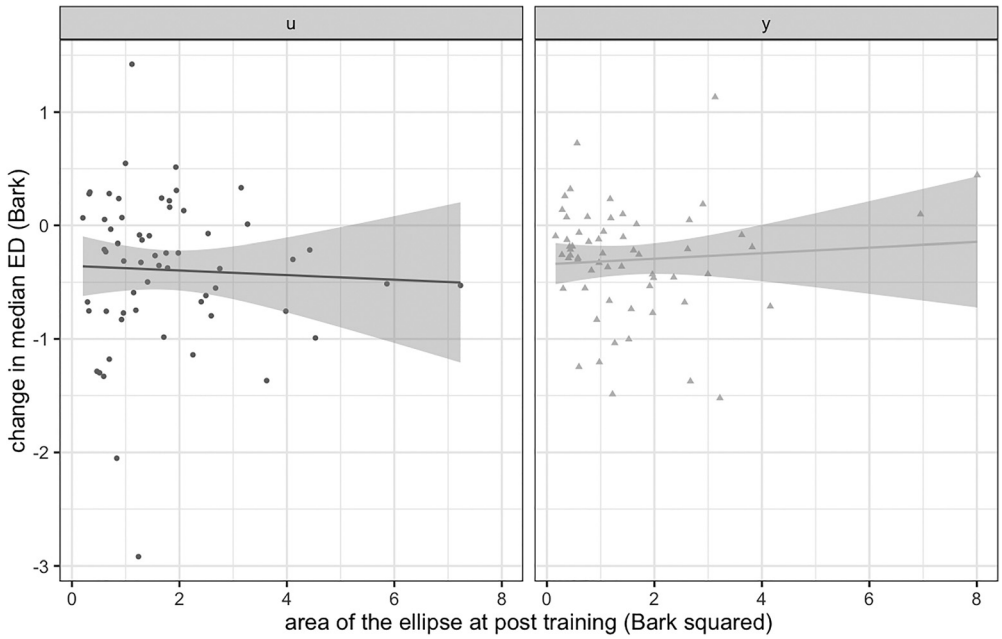

FIG. 8. Association between post-training variability (area of the ellipse) and change in median ED over the course of training, partitioned by vowel.

training condition that fits their individual needs. Therefore, the present study investigated several factors that could potentially predict outcomes in a biofeedback-enhanced L2 vowel learning task. Our hypothesized main predictors were auditory and somatosensory acuity, based on empirical evidence of associations between auditory perception and production accuracy in L2 learning (e.g., Flege, 1995; Bradlow *et al.*, 1997; Nagle, 2018; Hanulíková *et al.*, 2012), as well as theoretical and empirical evidence that speech is guided by somatosensory as well as auditory targets (Lametti *et al.*, 2012). Moreover, we hypothesized that the type of biofeedback provided—visual-acoustic or ultrasound—would interact with participants' profiles of acuity across sensory domains. Visual-acoustic biofeedback was hypothesized to provide a greater benefit to individuals with relatively low auditory acuity, since its clearly defined visual representation of the participant's acoustic output in relation to the target for a sound could compensate for less precise auditory perception. On the other hand, ultrasound biofeedback provides explicit visual information about articulator placement within the oral cavity, augmenting the tactile-kinesthetic feedback that is available under ordinary circumstances. Ultrasound was thus predicted to offer greater relative benefit to typical speakers with lower somatosensory acuity, who might otherwise have difficulty inferring what manipulations of the articulators are needed to achieve a particular output. Besides auditory and somatosensory acuity, we also tested the significance of other factors that have been found to predict learning outcomes in L2 production studies; these included production accuracy at baseline, phonological awareness, and production variability.

B. Predictors of baseline ED

Contrary to our hypothesis, performance on the AXB auditory discrimination task did not predict baseline production accuracy (ED relative to the native speaker target) for either vowel. Although this finding runs counter to some previous results (e.g., Bradlow *et al.*, 1997; Nagle, 2018; Hanulíková *et al.*, 2012), it aligns with other published literature that has reported no significant correlation between perceptual acuity and production accuracy in L2 learning (e.g., Kartushina and Frauenfelder, 2014; Peperkamp and Bouchon, 2011). Kartushina *et al.* (2015) did find a significant correlation between production variability and accuracy both before and after L2 production training, where lower variability was associated with higher accuracy. Other studies have suggested that such associations may be mediated by perceptual acuity (i.e., how narrowly target regions are specified in auditory-acoustic space), as well as language experience (Kartushina and Frauenfelder, 2014). However, in contrast with Kartushina *et al.* (2015), participants in the present study showed no significant association between variability and accuracy at baseline. Thus, our findings were indicative of dissociation between L2 perception and production abilities, at least in the early stage of acquisition represented by pre-training performance at the start of the present study.

As noted in Sec. III (Results), no individual predictors were significant in our analysis of baseline ED, and the regression model as a whole was non-significant, suggesting that other factors not considered in the present study will need to be measured to provide an explanation of variance in baseline accuracy. One notable oversight of the present study was our failure to collect information about participants' musicality and/or musical training, which have been found to predict L2 production accuracy in multiple studies (e.g., Christiner and Reiterer, 2015; Hu *et al.*, 2013; Milovanov *et al.*, 2010; Slevc and Miyake, 2006). Vocal training, in particular, has been associated with improved skill in imitating the sounds of an L2 (Christiner and Reiterer, 2015). We do note, however, that some literature has suggested that the

link between musical training and L2 production is indirect and could be mediated by differences in auditory and/or somatosensory acuity. For instance, Posedel *et al.* (2012) found that pitch perception ability was predictive of L2 production skill; musical training was significantly correlated with pitch perception ability but was not itself a significant predictor of production. Other literature has suggested that singing training may lead to enhanced somatosensory processing (Halwani *et al.*, 2011; Kleber and Zarate, 2014). Thus, while it will be important for future research to collect information about musical training as a predictor of L2 production, researchers should be aware of the potential for collinearity of this factor with direct measures of auditory and/or somatosensory acuity. In addition to musical training, future research should consider collecting measures of working memory (Reiterer *et al.*, 2011; Rota and Reiterer, 2009) and possibly other factors such as motivation (Dörnyei, 1998).

C. Predictors of change in ED

One of our primary experimental questions was whether participants would show a significant degree of change in L2 production accuracy after a brief duration (30 min per vowel) of biofeedback training. Our results showed a general decrease in ED from pre- to post-probe, demonstrating an overall learning effect associated with biofeedback training (Fig. 4). The magnitude of change was comparable to that reported in previous studies: our participants averaged 19.8%–23.6% change as a percentage of baseline ED, similar to the finding by Kartushina and colleagues (2016) of an 18%–20% change as a percentage of baseline MD. Kartushina *et al.* (2016) emphasized that such gains are particularly noteworthy when they occur in the context of an L2 target vowel with a close L1 counterpart, as was the case for Mandarin /u/ in the present study, since such targets are generally considered more difficult to acquire (Flege, 1995). Thus, the present findings add to previous literature reporting that biofeedback-enhanced training can produce measurable changes in L2 production accuracy even when targets are challenging and the duration of training is brief (Kartushina *et al.*, 2015; Gick *et al.*, 2008). Of course, the present design does not allow us to conclude that the observed changes are directly attributable to the inclusion of biofeedback in our training paradigm; a no-biofeedback comparison condition would be essential for such a claim. In addition, the regression examining change in ED yielded no significant main effect of biofeedback type (ultrasound versus visual-acoustic), but a follow-up equivalence test failed to reject the null hypothesis of equivalence across the two conditions. Thus, additional data collection will be needed to make claims about the efficacy of the types of biofeedback relative to non-biofeedback training and to one another. Finally, the present study examined outcomes for a highly limited task (imitation of isolated vowels) on a very short time frame. For this research to translate to clinical or pedagogical applications, it will be essential to conduct studies evaluating generalization to more naturalistic targets on a longer time frame.

For both vowels, accuracy in imitating the L2 target before training (baseline ED) was a significant predictor of change in ED from pre- to post-training, such that individuals with less accurate pre-training productions tended to show a greater degree of improvement in production over the course of training. This converges with other L2 pronunciation training studies that have reported that individuals who started out with poorer production accuracy made the greatest gains (Kartushina *et al.*, 2015; Bradlow *et al.*, 1997). The most straightforward interpretation of this finding is as a ceiling effect, where speakers who begin training with poorer productions have more room for improvement than individuals who begin with a higher degree of accuracy (Kartushina *et al.*, 2015).

The other two factors that were significantly associated with change in ED were found to behave asymmetrically across vowels: phonological awareness was a significant predictor of learning gains for /u/ but not /y/, while post-training variability was significantly associated with training outcomes for /y/ only. These asymmetries may relate to the distinction between Mandarin /u/ as an assimilated vowel with a close counterpart in the English vowel space (Flege, 1987) and Mandarin /y/ as an uncategorized vowel that lacks such a counterpart.

For the /u/ vowel only, phonological awareness (PACS) was a significant predictor of change in ED from pre- to post-training, with individuals with higher PACS scores tending to make greater progress over the course of training. This significant effect of phonological awareness agrees with previously reported results. Perrachione *et al.* (2011) found that, along with pre-training pitch perception ability, phonological awareness was a significant predictor of success in learning to perceive unfamiliar lexical tone contrasts. Although several other previous studies have also identified phonological awareness as a predictor of success in L2 learning (Hu, 2010; Hu, 2003; Hummel, 2009), only a few have examined the specific domain of L2 production accuracy (Nushi Kochaksaraie and Makiabadi, 2018; Hu *et al.*, 2013). Moreover, it remains somewhat unclear how phonological awareness has its impact on L2 learning, and this question is particularly understudied in the context of L2 production. Previous research has established that phonological awareness is related but not identical to perceptual acuity (McBride-Chang, 1996; Watson and Miller, 1993). However, in the present study, perceptual acuity was not a significant predictor of learning outcomes, and phonological awareness and perceptual acuity were not correlated. An alternative possibility is that learners with higher phonological awareness could be more attentive to small phonetic differences across languages (Mora *et al.*, 2014). Drawing on Schmidt's noticing hypothesis (Schmidt, 1993, 1990) which holds that awareness of an L2 exists along a continuum from unconsciously *perceiving* to consciously *noticing* to explicitly *understanding*, Mora *et al.* (2014) argued that awareness at least as high as the noticing level is needed for L2 phonological learning to occur. Applying these concepts to the present study, it is possible that individuals with higher phonological awareness, without necessarily differing in raw perceptual acuity, might be better able to notice and respond to the subtle acoustic cues that differentiate two similar

targets like /u/ in Mandarin and English. The absence of an association between phonological awareness and success in acquiring Mandarin /y/ is consistent with the fact that this vowel is unlikely to assimilate to an English vowel category, regardless of speakers' phonological awareness levels.

Last, for the /y/ vowel only, post-training production variability had a significant positive association with change in production accuracy—that is, speakers who were less variable at the post-training time point tended to have made greater progress over the course of training. This aligns with previous findings. Kartushina and Frauenfelder (2014) found that lower variability (in L1, as well as L2, production) was associated with higher accuracy in L2 production, and Kartushina *et al.* (2015) found that improvement over the course of training of non-native speech sounds was associated with a reduction in production variability. Low post-training production variability may be interpreted as an indication that speakers established a new speech sound category that they could use to consistently and accurately produce the non-native target (Kartushina *et al.*, 2016).

In contrast with the uncategorized vowel /y/, for /u/ there was no significant association between change in ED and production variability, either before or after training. This result differs from the findings of Kartushina and Frauenfelder (2014), who reported that the association between variability and accuracy in L2 production was stronger for assimilated than for uncategorized vowels in L2. However, our findings may be partially explained by another study conducted by Kartushina and colleagues (2016), which measured the effects of production training on accuracy and variability for two L2 vowels, one similar and one dissimilar to participants' L1 (French) categories. They found that the change in variability over the course of training was greater for the uncategorized L2 vowel than for the similar sound. On one interpretation, this smaller magnitude of change in variability could reflect the fact that some speakers had low variability prior to training because they were simply falling back on their motor plan for the similar L1 target, which they produce with high stability. In this view, the present finding that variability was not predictive of learning outcomes for the assimilated /u/ vowel may reflect the fact that speakers who showed low variability in producing /u/ can be subdivided into successful learners (who achieved a stable L2 target) and unsuccessful learners (who simply reused their pre-established articulatory pattern for English /u/ throughout the study).

Of course, some of the most noteworthy results of our analysis of change in ED pertain to the hypothesized predictors that were not found to be significant. Our measures of auditory and somatosensory acuity were not significantly associated with response to training, and we did not observe the hypothesized interaction between sensory acuity and biofeedback type. Possible explanations and interpretations are discussed in Sec. IV D.

D. Limitations and future directions

1. Limitations of our auditory acuity measure

Our analyses revealed that auditory acuity as measured in an AXB task was not predictive of either ED at baseline or change in ED, nor did we observe the hypothesized interaction between auditory acuity and biofeedback condition. However, as discussed above, we did find a significant association between token-to-token variability at the post-training time point and improvement over the course of training for the /y/ vowel, a result that aligns with a number of findings in previously published literature on perceptual acuity. Most notably, Kartushina and Frauenfelder (2014) found that production accuracy was not significantly associated with performance on a task explicitly measuring auditory acuity, but it did correlate significantly with token-to-token variability in production of both L1 and L2 vowels. Kartushina and Frauenfelder (2014) suggested that the observed association between variability and production accuracy could be explained by the fact that speakers with more compact vowel categories have more room in acoustic space to insert new nonnative categories. This is evocative of the DIVA framework and empirical work showing that individuals with more narrowly specified auditory targets tend to produce phonemic contrasts with lower variability and/or greater separation (Franken *et al.*, 2017; Perkell *et al.*, 2004a). This raises the possibility that production variability is simply a better index of perceptual acuity than explicit measures like the AXB discrimination task used in the present study. However, studies acknowledge that token-to-token variability also has a component of motor stability, which could, in principle, correlate with somatosensory acuity (Franken *et al.*, 2017; Kartushina *et al.*, 2016). In our data, we found no correlation between baseline variability and either auditory or somatosensory acuity.

In total, the present findings suggest that further research is warranted to better understand the factors that drive individual differences in production variability and the impact of these differences on learning outcomes. In future research we intend to obtain more comprehensive measures of production variability, including both L1 and L2 productions. In addition, our variability measure is somewhat limited in that we examined variability only at the midpoint of the vowel, while some participants showed fluctuating formant frequencies from the beginning to the end of their vowel productions. A recent body of work by Niziolek and colleagues (e.g., Niziolek *et al.*, 2013) has emphasized the dynamic nature of variability over the course of individual utterances, with productions that start out furthest from a speaker's average formant frequencies for a given sound tending to exhibit "centering" by mid-vowel. Follow-up analyses that examine variability at onset and change in variability from onset to mid-vowel could help elucidate the relationship between production variability and L2 pronunciation skill.

An additional limitation of the auditory perceptual task we used pertains to the fact that we had to choose to focus on a specific part of the synthetic continuum in order to calculate JND; that is, we had to choose to start either at the /y/ end, the /u/ end, or at the boundary between the phonemes. We decided that starting from an endpoint was more consistent with the categorical nature of speech perception in normal circumstances, and we arbitrarily opted to start from the /y/ end. As a consequence, though, our auditory acuity measure was more relevant to the vowel /y/ than /u/. Since the

assimilated /u/ vowel turned out to be more challenging to acquire than uncategorized /y/, a more nuanced measure of perception at the /u/ end of the continuum could prove illuminating. In the future, running the AXB task from both ends of the continuum and calculating an average JND might represent the most effective strategy.

2. Limitations of our somatosensory acuity measure

Contrary to hypothesis, somatosensory acuity was not found to interact with biofeedback type in predicting change in ED. However, there are major limitations to our somatosensory measure in the context of the vowel-learning task that formed the focus of the current study. Previous studies that utilized a spatial resolution task to measure somatosensory acuity (Ghosh *et al.*, 2010; Perkell *et al.*, 2004b) measured participants' production of the consonants /s/ and /ʃ/ rather than vowels. The tighter constriction of the vocal tract results in stronger oral-tactile feedback for consonants than vowels, and somatosensory targets may play a correspondingly greater role (Ghosh *et al.*, 2010; Guenther, 2016). In addition, our stereognosis task measured the tactile acuity of the tongue tip, which is potentially relevant for coronal consonants like /s/ and /ʃ/ but less applicable to vowels, where the tongue dorsum is the active articulator. We used the oral stereognosis task in spite of these limitations for two primary reasons. First, previous research has reported group differences in similar oral stereognosis tasks between individuals with and without misarticulation of sounds that do not necessarily involve tactile feedback to the tongue tip, notably /ɹ/ (Fucci, 1972; Fucci and Robertson, 1971; McNutt, 1977). Second, the existing literature offers very few options for somatosensory measurement targeted to the properties of vowels. One exception is Zandipour *et al.* (2006), who piloted a task in which speakers were asked to produce vowels while compensating for a somatosensory perturbation (a bite block) in the context of simultaneous auditory masking. Without auditory feedback, speakers must rely on oral somatosensation to adapt their speech to the presence of the bite block. It remains possible that, by using a somatosensory measure more directly relevant to the phoneme targets of interest, we might find evidence for the hypothesized interaction between somatosensory acuity and response to different types of biofeedback.

A final limitation of the somatosensory acuity task used here pertains to the fact that, following the protocol from Steele *et al.* (2014), the same 7 letters were used for each step in the adaptive staircase; this contrasts with previous studies of oral stereognosis, which have used as many as 20 different forms (Fucci and Robertson, 1971). Although we did not see clear evidence of ceiling-level performance on participants' first attempts, numerous participants achieved perfect performance on the second run, suggesting that they may have developed heuristics pertaining to the limited set of possible responses. In sum, even if our experimental learning task involved more relevant targets like coronal consonants, it may still be desirable to use a more challenging task than that from Steele *et al.* (2014).

3. Sensory asymmetry versus sensory acuity

Last, in our initial framing of the hypothesized interaction between biofeedback type and sensory acuity in auditory and somatosensory domains, we were implicitly assuming some degree of asymmetry across domains. That is, our hypotheses specifically apply to individuals with low acuity in one domain and intact sensation in the other domain. For instance, we posited that an individual with low somatosensory acuity in the context of typical auditory acuity would show a greater response to ultrasound than visual-acoustic biofeedback. This reasoning does not generate a prediction for a preferred form of biofeedback for individuals who are either strong or weak in both domains. Therefore, it is possible that it would be better to frame our predictions in terms of sensory asymmetry rather than raw sensory acuity. In an exploratory follow-up analysis, we found no difference in outcomes when we examined the interaction between biofeedback type and difference in normalized sensory acuity (somatosensory - auditory) instead of individual acuity scores. However, there was a numerical trend in the predicted direction. Among individuals with asymmetric acuity favoring the somatosensory domain, those who received visual-acoustic biofeedback tended to show a greater change in ED than those who received ultrasound biofeedback. Among individuals with an asymmetry in the other direction, change in ED was greater for those who received ultrasound than those who received visual-acoustic biofeedback. In addition, more than half of the participants in the present sample had less than one unit of difference between standardized acuity scores in auditory and somatosensory domains. It remains possible that significant differences in response to different biofeedback types could emerge if we were to use a sample specifically selected to feature a high degree of sensory asymmetry. In future research, it could be particularly illuminating to conduct within-subject comparisons of visual-acoustic and ultrasound biofeedback in individuals hand-selected for asymmetric acuity across auditory and somatosensory domains.

V. CONCLUSION

This study investigated possible predictors of response to ultrasound and visual-acoustic biofeedback training in an L2 learning task in which native English speakers were taught to produce the Mandarin vowels /u/ and /y/. On average, participants improved their production, as indicated by a decrease in ED relative to a native-speaker target, with moderate effect size across both vowels and biofeedback conditions. The data reported here did not provide evidence supporting our hypothesis of an interaction between sensory acuity and biofeedback type, such that learners with lower acuity in a given sensory domain would show a greater magnitude of response to the biofeedback modality that targets that domain. Future research should test if refined measures of auditory and somatosensory acuity, or measures that specifically quantify asymmetry between these domains, might more directly predict response to different types of biofeedback training. Despite the null result in our primary question of interest, the present study yielded several findings that can

inform future research and clinical or pedagogical practice. First, we extended a small existing literature with our finding that higher phonological awareness was associated with improved learning outcomes for production of an L2 vowel target with a close counterpart in L1. Such results have the potential to inform pedagogical practice, suggesting that language teachers may wish to make phonological awareness an explicit part of L2 pronunciation training, particularly when dealing with speech sounds that have similar counterparts in the learners' other language(s). Second, we found that better learning outcomes were associated with reduced production variability at the post-training time point, a result that was again mediated by vowel. This adds to a growing body of literature suggesting that variability across repeated productions may be an important measure for understanding sensorimotor control of speech. Although considerable work remains to be done, findings from this line of research could ultimately inform pedagogical and clinical decisions when selecting the optimal training paradigm for a given learner, with the long-term goal of maximizing learning in speech production training tasks.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (Grant Nos. NIDCD-R01DC013668 and NIDCD-R01DC017476). The authors gratefully thank the following individuals: Daniel Lametti and Mark Tiede for assistance with auditory stimulus generation, members of the Biofeedback Intervention Technology for Speech Lab (BITS Lab) at New York University (NYU) for assistance with data collection and analysis, and all participants for their time.

[1]Other current models, such as the Perception for Action Control Theory (Schwartz *et al.*, 2012) or the Hierarchical State Feedback Control Model (Hickok, 2012), could provide an equally suitable theoretical framework for our approach; however, we focus on the DIVA model given that it has been computationally implemented.

[2]This hypothesis was substantiated by measuring token-to-token variability across repeated productions of English /u/, as well as Mandarin /u/, in a subset of 16 participants from the present study. After excluding one subject whose measure of area of the ellipse for English /u/ was an extreme outlier, there was a significant correlation of moderate magnitude between variability in English and Mandarin, $r(13) = 0.66$, $p < 0.01$. We do not have this measure for all participants because, unfortunately, we only realized the importance of obtaining this information after data collection was already underway.

[3]The decision to include only female speakers was intended to minimize variability in vocal tract size, which is relevant in a task of matching formants or vocal tract configurations. However, this experimental choice raises the possibility that results from this study may only be generalizable to female speakers.

[4]Regional dialect and language background data are not being reported at the individual subject level because such information can, in some cases, be individually identifying.

[5]Several participants who achieved ceiling-level accuracy on the second run of the task mentioned that they had become attuned to a slight artifact that differentiated the standard stimulus (X in the AXB task) from the other stimuli in the continuum. This supports the idea that their ceiling-level performance was attributable to learning of the task rather than an improvement in perceptual acuity from the first to the second task run.

[6]The Mandarin speakers were selected to represent different heights (>170 cm, 150–170 cm, and <150 cm, denoted as "tall," "medium," and "short," respectively, in the data file control_data.csv, located at https://osf.io/9djpm/), since height correlates with vocal tract length and, thus, with average formant frequencies (Fitch, 1997; Fitch and Giedd, 1999). The speakers came from different regions of origin (Taiyuan, Guilin, and Wuhan, respectively), but there were no discernable dialectal influences on their productions of the vowels in question.

[7]Participants held the probes in place themselves; no stabilizing device was used, but the experimenter helped the participant re-position the probe if it was observed to move away from the midline.

[8]We additionally examined Mahalanobis distance (MD) as an index of production accuracy. MD, which has been used in similar previous research (Kartushina *et al.*, 2015), quantifies how many standard deviations (SDs) away from the target sound distribution in F1-F2 space a production falls, where SDs are multidimensional and defined by the principal axes of the target distribution. However, visual inspection of individually plotted data showed occasional cases of mismatch between computed MD values and visual gestalt impressions (e.g., cases where MD increased from pre- to post-training, but the plotted data suggested that the participant had gotten closer to the native speaker target). This may reflect the fact that our native speaker sample was quite small (six repetitions of each vowel by three speakers) and, therefore, could not be assumed to provide an accurate estimate of the variability intrinsic to the distribution of tokens in the true population.

[9]For the first vowel trained, the midpoint probe was the immediate post-training probe; for the second vowel, the final probe of the session was the immediate post-training probe.

[10]Kartushina and colleagues use the term "compactness" to describe token-to-token variability in phonetic production, where a tighter distribution has a higher compactness. However, the actual measure that they compute is the area of an ellipse encompassing the distribution of productions, and a larger area corresponds with a more diffuse, less compact distribution. To avoid this incongruity between terminology and computation, we favor the term "variability" over compactness in the present paper.

[11]Nonparametric measures were used because not all participants' productions were normally distributed.

[12]Although individual subjects' ED values were not necessarily normally distributed and thus were summarized using medians, the group-level distribution of median ED values was normal (Shapiro-Wilk, $p = 0.89$) and was therefore summarized using mean and SD.

Adler-Bock, M., Bernhardt, B. M., Gick, B., and Bacsfalvi, P. (**2007**). "The use of ultrasound in remediation of North American English /r/ in 2 adolescents," Am. J. Speech-Lang. Pathol. **16**(2), 128–139.

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., and Pruitt, J. S. (**1998**). "Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores," paper presented at the *Fifth International Conference on Spoken Language Processing*.

Bacsfalvi, P., and Bernhardt, B. M. (**2011**). "Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography," Clin. Linguist. Phonet. **25**(11-12), 1034–1043.

Bernhardt, B., Gick, B., Bacsfalvi, P., and Adler-Bock, M. (**2005**). "Ultrasound in speech therapy with adolescents and adults," Clin. Linguist. Phonet. **19**(6-7), 605–617.

Best, C. T. (**1995**). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Baltimore, MD), pp. 171–204.

Best, C. T., Halle, P., Bohn, O.-S., and Faber, A. (**2003**). "Cross-language perception of nonnative vowels: Phonological and phonetic effects of listeners' native languages," paper presented at the *Proceedings of the 15th International Congress of Phonetic Sciences*.

Bliss, H., Abel, J., and Gick, B. (**2018**). "Computer-assisted visual articulation feedback in L2 pronunciation instruction," J. Second Lang. Pronunciation **4**(1), 129–153.

Boersma, P., and Weenink, D. (**2016**). "Praat: Doing phonetics by computer (version 6.0.18) [computer program]," http://www.praat.org (Last viewed 5 April 2019).

Borrie, S. A., and Schäfer, M. C. (**2015**). "The role of somatosensory information in speech perception: Imitation improves recognition of disordered speech," J. Speech Lang. Hear. Res. **58**(6), 1708–1716.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. I. (**1997**). "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**(4), 2299–2310.

Carey, M. (**2004**). "CALL visual feedback for pronunciation of vowels: Kay Sona-Match," CALICO J. **21**, 571–601.
Catford, J. C., and Pisoni, D. (**1970**). "Auditory versus articulatory training in exotic sounds," Mod. Lang. J. **54**, 477–481.
Chang, C. B., Yao, Y., Haynes, E. F., and Rhodes, R. (**2011**). "Production of phonetic and phonological contrast by heritage speakers of Mandarin," J. Acoust. Soc. Am. **129**(6), 3964–3980.
Chen, Y., Robb, M., Gilbert, H., and Lerman, J. (**2001**). "Vowel production by Mandarin speakers of English," Clin. Linguist. Phonet. **15**(6), 427–440.
Christiner, M., and Reiterer, S. M. (**2015**). "A Mozart is not a Pavarotti: Singers outperform instrumentalists on foreign accent imitation," Front. Hum. Neurosci. **9**, 482.
Dörnyei, Z. (**1998**). "Motivation in second and foreign language learning," Lang. Teach. **31**(3), 117–135.
Dowd, A., Smith, J., and Wolfe, J. (**1998**). "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," Lang. Speech **41**(1), 1–20.
Engwall, O. (**2006**). "Feedback strategies of human and virtual tutors in pronunciation training," Speech Music and Hearing—Quarterly Progress and Status Report Vol. **48**(1), 11–34.
Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," J. Acoust. Soc. Am. **102**(2), 1213–1222.
Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**(3), 1511–1522.
Flege, J. E. (**1987**). "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," J. Phonetics **15**(1), 47–65.
Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience*, edited by W. Strange (York Press, Timonium, MD), pp. 233–277.
Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., and Hagoort, P. (**2017**). "Individual variability as a window on production-perception interactions in speech motor control," J. Acoust. Soc. Am. **142**(4), 2007–2018.
Fucci, D. (**1972**). "Oral vibrotactile sensation: An evaluation of normal and defective speakers," J. Speech Lang. Hear. Res. **15**(1), 179–184.
Fucci, D. J., and Robertson, J. H. (**1971**). "'Functional' defective articulation: An oral sensory disturbance," Percept. Mot. Skills **33**(3), 711–714.
Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Menard, L., Guenther, F. H., Lane, H., and Perkell, J. S. (**2010**). "An investigation of the relation between sibilant production and somatosensory and auditory acuity," J. Acoust. Soc. Am.**128**(5), 3079–3087.
Gick, B., Bernhardt, B., Bacsfalvi, P., Wilson, I., and Zampini, M. (**2008**). "Ultrasound imaging applications in second language acquisition," Phonol. Second Lang. Acquis. **36**, 315–328.
Gilbert, J. B. (**2010**). "Pronunciation as orphan: What can be done?," Speak Out! **43**, 3–7.
Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Biol. Cybernet. **72**(1), 43–53.
Guenther, F. H. (**1995**). "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," Psychol. Rev. **102**(3), 594.
Guenther, F. H. (**2016**). *Neural Control of Speech* (MIT Press, Cambridge, MA).
Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain Lang. **96**(3), 280–301.
Guenther, F. H., Hampson, M., and Johnson, D. (**1998**). "A theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. **105**(4), 611.
Halwani, G. F., Loui, P., Rueber, T., and Schlaug, G. (**2011**). "Effects of practice and experience on the arcuate fasciculus: Comparing singers, instrumentalists, and non-musicians," Front. Psychol. **2**, 156.
Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., and Huettig, F. (**2012**). "Individual differences in the acquisition of a complex L2 phonology: A training study," Lang. Learn. **62**, 79–109.
Hardcastle, W. J., Gibbon, F. E., and Jones, W. (**1991**). "Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders," Br. J. Disord. Commun. **26**(1), 41–74.
Hattori, K., and Iverson, P. (**2010**). "Examination of the relationship between L2 perception and production: An investigation of English /r/-/l/ perception and production by adult Japanese speakers," paper presented at the *Second Language Studies: Acquisition, Learning, Education and Technology*.
Hickok, G. (**2012**). "Computational neuroanatomy of speech production," Nat. Rev. Neurosci. **13**(2), 135.
Hitchcock, E. R., McAllister Byun, T., Swartz, M., and Lazarus, R. (**2017**). "Efficacy of electropalatography for treating misarticulation of /r/," Am. J. Speech-Lang. Pathol. **26**(4), 1141–1158.
Hu, C. F. (**2003**). "Phonological memory, phonological awareness, and foreign language word learning," Lang. Learn. **53**(3), 429–462.
Hu, C. F. (**2010**). "Phonological bases for L2 morphological learning," J. Psycholinguist. Res. **39**(4), 305–322.
Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., and Reiterer, S. M. (**2013**). "Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates," Brain Lang. **127**(3), 366–376.
Hummel, K. M. (**2009**). "Aptitude, phonological memory, and second language proficiency in nonnovice adult learner," Appl. Psycholinguist. **30**(2), 225–249.
Jacobs, R., Serhal, C. B., and van Steenberghe, D. (**1998**). "Oral stereognosis: A review of the literature," Clin. Oral Investig. **2**(1), 3–10.
Kartushina, N., and Frauenfelder, U. H. (**2014**). "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," Front. Psychol. **5**, 1246.
Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. (**2015**). "The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds," J. Acoust. Soc. Am. **138**(2), 817–832.
Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. (**2016**). "Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training," J. Phonetics **57**, 21–39.
Kawahara, H., Morise, M., Banno, H., and Skuk, V. G. (**2013**). "Temporally variable multi-aspect *N*-way morphing based on interference-free speech representations," paper presented at the *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Asia-Pacific.
Kendall, T., Thomas, E. R., and Kendall, M. T. (**2018**). "vowels: Vowel manipulation, normalization, and plotting [*R* package]," available at https://cran.r-project.org/web/packages/vowels/ (Last viewed 5 April 2019).
Kleber, B. A., and Zarate, J. M. (**2014**). "The neuroscience of singing," in *The Oxford Handbook of Singing* (Oxford University Press, Oxford, UK).
Kocjančičo Antolík, T. K., Pillot-Loiseau, C., and Kamiyama, T. (**2019**). "The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training," J. Second Lang. Pronunciation **5**(1), 72–97.
Lakens, D. (**2017**). "Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses," Soc. Psychol. Personal. Sci. **8**(4), 355–362.
Lametti, D. R., Nasir, S. M., and Ostry, D. J. (**2012**). "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback," J. Neurosci. **32**(27), 9351–9358.
Lennes, M. (**2003**). "Collect_formant_data_from_files, Praat [Praat script]," available at https://github.com/FieldDB/Praat-Scripts/blob/master/collect_formant_data_from_files.praat (Last viewed 5 April 2019).
Levy, E. S., and Strange, W. (**2008**). "Perception of French vowels by American English adults with and without French language experience," J. Phonetics **36**(1), 141–157.
McAllister Byun, T. (**2017**). "Efficacy of visual–acoustic biofeedback intervention for residual rhotic errors: A single-subject randomization study," J. Speech Lang. Hear. Res. **60**(5), 1175–1193.
McAllister Byun, T., and Campbell, H. (**2016**). "Differential effects of visual-acoustic biofeedback intervention for residual speech errors," Front. Hum. Neurosci. **10**, 567.
McAllister Byun, T., and Hitchcock, E. R. (**2012**). "Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation," Am. J. Speech-Lang. Pathol. **21**(3), 207–221.
McAllister Byun, T., Hitchcock, E. R., and Swartz, M. T. (**2014**). "Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention," J. Speech Lang. Hear. Res. **57**(6), 2116–2130.
McBride-Chang, C. (**1996**). "Models of speech perception and phonological processing in reading," Child Dev. **67**(4), 1836–1856.
McNutt, J. C. (**1977**). "Oral sensory and motor behaviors of children with /s/ or /r/ misarticulations," J. Speech Lang. Hear. Res. **20**(4), 694–703.

Milovanov, R., Pietilä, P., Tervaniemi, M., and Esquef, P. A. (**2010**). "Foreign language pronunciation skills and musical aptitude: A study of Finnish adults with higher education," Learn. Individ. Differ. **20**(1), 56–60.

Mora, J. C., Rochdi, Y., and Kivistö-de Souza, H. (**2014**). "Mimicking accented speech as L2 phonological awareness," Lang. Aware. **23**(1-2), 57–75.

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

Nasir, S. M., and Ostry, D. J. (**2008**). "Speech motor learning in profoundly deaf adults," Nat. Neurosci. **11**(10), 1217–1222.

Newman, R. S. (**2003**). "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," J. Acoust. Soc. Am. **113**(5), 2850–2860.

Niziolek, C. A., Nagarajan, S. S., and Houde, J. F. (**2013**). "What does motor efference copy represent? Evidence from speech production," J. Neurosci. **33**(41), 16110–16116.

Nushi Kochaksaraie, M., and Makiabadi, H. (**2018**). "Second language learners' phonological awareness and perception of foreign accentedness and comprehensibility by native and non-native English speaking EFL teachers," J. Teach. Lang. Skills **36**(4), 103–140.

Okuno, T., and Hardison, D. M. (**2016**). "Perception-production link in L2 Japanese vowel duration: Training with technology," Lang. Learn. Technol. **20**(2), 61–80.

Olson, D. J. (**2014**). "Benefits of visual feedback on segmental production in the L2 classroom," Lang. Learn. Technol. **18**(3), 173–192.

Peperkamp, S., and Bouchon, C. (**2011**). "The relation between perception and production in L2 phonological processing," paper presented at the *Twelfth Annual Conference of the International Speech Communication Association*.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., and Zandipour, M. (**2004a**). "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," J. Acoust. Soc. Am. **116**(4), 2338–2344.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., and Guenther, F. H. (**2004b**). "The distinctness of speakers' /s/-/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect," J. Speech Lang. Hear. Res. **47**(6), 1259–1269.

Perrachione, T. K., Lee, J., Ha, L. Y., and Wong, P. C. (**2011**). "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," J. Acoust. Soc. Am. **130**(1), 461–472.

Posedel, J., Emery, L., Souza, B., and Fountain, C. (**2012**). "Pitch perception, working memory, and second-language phonological production," Psychol. Music **40**(4), 508–517.

Preston, J. L., and Leaman, M. (**2014**). "Ultrasound visual feedback for acquired apraxia of speech: A case report," Aphasiology **28**(3), 278–295.

Preston, J. L., Leece, M. C., and Maas, E. (**2016**). "Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia," Front. Hum. Neurosci. **10**, 440.

Preston, J. L., McAllister, T., Phillips, E., Boyce, S., Tiede, M., Kim, J. S., and Whalen, D. H. (**2018**). "Treatment for residual rhotic errors with high- and low-frequency ultrasound visual feedback: A single-case experimental design," J. Speech Lang. Hear. Res. **61**(8), 1875–1892.

Preston, J. L., McAllister, T., Phillips, E., Boyce, S. E., Tiede, M., Kim, J. S., and Whalen, D. H. (**2019**). "Remediating residual rhotic errors with traditional and ultrasound-enhanced treatment: A single-case experimental study," Am. J. Speech-Lang. Pathol. **28**, 1167–1183.

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., and Maas, E. (**2014**). "Ultrasound visual feedback treatment and practice variability for residual speech sound errors," J. Speech Lang. Hear. Res. **57**(6), 2102–2115.

R Core Team. (**2015**). "*R*: A language and environment for statistical computing [software]," (R foundation for Statistical Computing, Vienna, Austria), available at http://www.R-project.org/ (Last viewed 5 April 2019).

Reiterer, S. M., Hu, X., Erb, M., Rota, G., Nardo, D., Grodd, W., and Ackermann, H. (**2011**). "Individual differences in audio-vocal speech imitation aptitude in late bilinguals: Functional neuro-imaging and brain morphology," Front. Psychol. **2**, 271.

Rota, G., and Reiterer, S. (**2009**). "Cognitive aspects of pronunciation talent," Lang. Talent Brain Act. **1**, 67–96.

Schmidt, R. (**1990**). "The role of consciousness in second language learning," Appl. Linguist. **11**(2), 129–158.

Schmidt, R. (**1993**). "Awareness and second language acquisition," Annu. Rev. Appl. Linguist. **13**, 206–226.

Schwartz, J.-L., Basirat, A., Ménard, L., and Sato, M. (**2012**). "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," J. Neurolinguist. **25**(5), 336–354.

Simmonds, A. J., Wise, R. J., Dhanjal, N. S., and Leech, R. (**2011**). "A comparison of sensory-motor activity during speech in first and second languages," J. Neurophysiol. **106**(1), 470–478.

Slevc, L. R., and Miyake, A. (**2006**). "Individual differences in second-language proficiency: Does musical ability matter?," Psychol. Sci. **17**(8), 675–681.

Steele, C. M., Hill, L., Stokely, S., and Peladeau-Pigeon, M. (**2014**). "Age and strength influences on lingual tactile acuity," J. Texture Stud. **45**(4), 317–323.

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. **122**(4), 2306–2319.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. R. (**2013**). *Comprehensive Test of Phonological Processing*, Second ed. (Pro-Ed, Austin, TX).

Watson, B. U., and Miller, T. (**1993**). "Auditory perception, phonological processing, and reading ability/disability," J. Speech Lang. Hear. Res. **36**(4), 850–863.

Wickham, H. (**2009**). "ggplot2: Elegant graphics for data analysis [*R* package]," available at https://doi.org/http://ggplot2.org (Last viewed 5 April 2019).

Wickham, H. (**2016**). "tidyr: Easily tidy data with 'spread()' and 'gather()' functions [*R* package]," available at https://doi.org/http://cran.r-project.org/package=tidy (Last viewed 5 April 2019).

Wickham, H., and Francois, R. (**2015**). "dplyr: A grammar of data manipulation [*R* package]," available at https://doi.org/http://CRAN.R-project.org/package=dplyr (Last viewed 5 April 2019).

Wong, J. W. S. (**2013**). "The effects of perceptual and/or productive training on the perception and production of English vowels /ɪ/ and /iː/ by Cantonese ESL learners," in *INTERSPEECH*, pp. 2113–2117.

Wong, P. C., and Perrachione, T. K. (**2007**). "Learning pitch patterns in lexical identification by native English-speaking adults," Appl. Psycholinguist. **28**(04), 565–585.

Wong, P. C., Vuong, L. C., and Liu, K. (**2017**). "Personalized learning: From neurogenetics of behaviors to designing optimal language training," Neuropsychologia **98**, 192–200.

Zandipour, M., Perkell, J., Guenther, F., Tiede, M., Honda, K., and Murano, E. (**2006**). "Speaking with a bite-block: Data and modeling," paper presented at the *Proceedings of the 7th International Seminar on Speech Production*.