frontiers
in Human Neuroscience

ORIGINAL RESEARCH
published: 19 November 2015
doi: 10.3389/fnhum.2015.00612

Visual Feedback of Tongue Movement for Novel Speech Sound Learning

William F. Katz* and Sonya Mehta

Speech Production Lab, Callier Center for Communication Disorders, School of Behavioral and Brain Sciences, The University of Texas at Dallas, Dallas, TX, USA

Pronunciation training studies have yielded important information concerning the processing of audiovisual (AV) information. Second language (L2) learners show increased reliance on bottom-up, multimodal input for speech perception (compared to monolingual individuals). However, little is known about the role of viewing one's own speech articulation processes during speech training. The current study investigated whether real-time, visual feedback for tongue movement can improve a speaker's learning of non-native speech sounds. An interactive 3D tongue visualization system based on electromagnetic articulography (EMA) was used in a speech training experiment. Native speakers of American English produced a novel speech sound (/ᶁ̠/; a voiced, coronal, palatal stop) before, during, and after trials in which they viewed their own speech movements using the 3D model. Talkers' productions were evaluated using kinematic (tongue-tip spatial positioning) and acoustic (burst spectra) measures. The results indicated a rapid gain in accuracy associated with visual feedback training. The findings are discussed with respect to neural models for multimodal speech processing.

Keywords: speech production, second language learning, visual feedback, audiovisual integration, electromagnetic articulography, articulation therapy

OPEN ACCESS

Edited by:
Marcelo L. Berthier,
University of Malaga, Spain

Reviewed by:
Peter Sörös,
University of Western Ontario, Canada
Caroline A. Niziolek,
Boston University, USA

****Correspondence:***
William F. Katz
wkatz@utdallas.edu

Received: *13 June 2015*
Accepted: *26 October 2015*
Published: *19 November 2015*

Citation:
Katz WF and Mehta S (2015) Visual Feedback of Tongue Movement for Novel Speech Sound Learning. Front. Hum. Neurosci. 9:612. doi: 10.3389/fnhum.2015.00612

INTRODUCTION

Natural conversation is a multimodal process, where the visual information contained in a speaker's face plays an important role in decoding the speech signal. Integration of the auditory and visual modalities has long been known to be more advantageous to speech perception than either input alone. Early studies of lip-reading found that individuals with hearing loss could more accurately recognize familiar utterances when provided with both auditory and visual cues compared to either modality on its own (Numbers and Hudgins, 1948; Erber, 1975). Research on healthy hearing populations has also shown that audiovisual integration enhances comprehension of spoken stimuli, particularly in noisy environments or situations where the speaker has a strong foreign accent (O'Neill, 1954; Sumby and Pollack, 1954; Erber, 1975; Reisberg et al., 1987). Even under optimal listening conditions, observing a talker's face improves comprehension for complex utterances, suggesting that visual correlates of speech movement are a central component to processing speech sounds (Reisberg et al., 1987; Arnold and Hill, 2001).

Studies investigating how listeners process conflicting audio and visual signals also support a critical role of the visual system during speech perception (McGurk and MacDonald, 1976; Massaro, 1984; Summerfield and McGrath, 1984). For example, listeners presented with the auditory signal for "ba" concurrently with the visual signal for "ga" typically report a blended percept, the well-known "McGurk effect." A recent study by Sams et al. (2005) demonstrated that the McGurk effect occurs even if the source of the visual input is the listener's *own* face.

In this study, subjects wore headphones and silently articulated a “pa” or “ka” while observing their productions in a mirror as a congruent or incongruent audio stimulus was simultaneously presented. In addition to replicating the basic McGurk (blended) effect, researchers found that simultaneous silent articulation alone moderately improved auditory comprehension, suggesting that knowledge from one’s own motor experience in speech production is also exploited during speech perception. Other cross-modal studies support this view. For instance, silently articulating a syllable in synchrony with the presentation of a concordant auditory and/or visually ambiguous speech stimulus has been found to improve syllable identification, with concurrent mouthing further speeding the perceptual processing of a concordant stimulus (Sato et al., 2013; also see Mochida et al., 2013; D’Ausilio et al., 2014). Taken together, these studies indicate that listeners benefit from multimodal speech information during the perception process.

Audiovisual (AV) information also plays an important role in acquiring novel speech sounds, according to studies of second language (L2) learning. Research has shown that speech comprehension by non-native speakers is influenced by the presence/absence of visual input (see Marian, 2009, for review). For instance, Spanish-speakers exposed to Catalan can better discriminate the non-native tense-lax vowel pair /e/ and /ε/ when visual information is added (Navarra and Soto-Faraco, 2007).

Computer-assisted pronunciation training (CAPT) systems have provided a new means of examining AV processing during language learning. Many CAPT systems, such as “Baldi” (Massaro and Cohen, 1998; Massaro, 2003; Massaro et al., 2006), “ARTUR” (Engwall et al., 2006; Engwall and Bälter, 2007; Engwall, 2008), “ATH” (Badin et al., 2008), “Vivian” (Fagel and Madany, 2008), and “Speech Tutor” (Kröger et al., 2010), employ animated talking heads, most of which can optionally display transparent vocal tracts showing tongue movement. “Tongue reading” studies based on these systems have shown small but consistent perceptual improvement when tongue movement information is added to the visual display. Such effects have been noted in word retrieval for acoustically degraded sentences (Wik and Engwall, 2008) and in a forced-choice consonant identification task (Badin et al., 2010).

Whereas the visual effects on speech perception are fairly well-established, the visual effects on speech production are less clearly understood. Massaro and Light (2003) investigated the effectiveness of using Baldi in teaching non-native phonetic contrasts (/r/-/l/) to Japanese learners of English. Both external and internal views (i.e., showing images of the speech articulators) of Baldi were found to be effective, with no added benefit noted for the internal articulatory view. A subsequent, rather preliminary report on English-speaking students learning Chinese and Arabic phonetic contrasts reported similar negative results for the addition of visual, articulatory information (Massaro et al., 2008). In this study, training with the Baldi avatar showing face (Mandarin) or internal articulatory processes (Arabic) provided no significant improvement in a small group of students’ productions, as rated by native listeners.

In contrast, Liu et al. (2007) observed potentially positive effects of visual feedback on speech production for 101 English-speaking students learning Mandarin. This investigation contrasted three feedback conditions: audio only, human audiovisual, and a Baldi avatar showing visible articulators. Results indicated that all three methods improved students’ pronunciation accuracy. However, for the final rime pronunciation both the human audiovisual and Baldi condition scores were higher than audio-only, with the Baldi condition significantly higher than the audio condition. This pattern is compatible with the view that information concerning the internal articulators helps relay information to assist in L2 production. Taken together, these studies suggest that adding visual articulatory information to 3D tutors can lead to improvements for producing certain language contrasts. However, more work is needed to establish the effectiveness, consistency, and strength of these techniques.

At the neurophysiological level, AV speech processing can be related to the issue of whether speech perception and production is supported by a joined action-observation matching system. Such a system has been related to “mirror” neurons originally described in the macaque brain [for reviews see (Rizzolatti and Craighero, 2004; Pulvermüller and Fadiga, 2010; Rizzolatti et al., 2014); although see (Hickok, 2009, 2010) for an opposing view]. Mirror neurons are thought to fire both during goal-directed actions and while watching a similar action made by another individual. Research has extended this finding to audiovisual systems in monkeys (Kohler et al., 2002) and speech processing in humans (e.g., Rizzolatti and Arbib, 1998; Arbib, 2005; Gentilucci and Corballis, 2006).

In support of this view, studies have linked auditory and/or visual speech perception with increased activity in brain areas involved in motor speech planning, execution, and proprioceptive control of the mouth (e.g., Möttönen et al., 2004; Wilson et al., 2004; Ojanen et al., 2005; Skipper et al., 2005, 2006, 2007a,b; Pekkola et al., 2006; Pulvermüller et al., 2006; Wilson and Iacoboni, 2006; Zaehle et al., 2008). Similarly, magnetoencephalography (MEG) studies have linked speech production with activity in brain areas specialized for auditory and/or visual speech perception processes (e.g., Curio et al., 2000; Gunji et al., 2001; Houde et al., 2002; Heinks-Maldonado et al., 2006; Tian and Poeppel, 2010). While auditory activation during speech production is expected (because acoustic input is normally present), Tian and Poeppel’s (2010) study shows auditory cortex activation in the absence of auditory input. This suggests that an imaginary motor speech task can nevertheless generate forward predictions via an auditory efference copy.

Overall, these neurophysiological findings suggest a brain basis for the learning of speech motor patterns via visual input, which in turn would strengthen the multimodal speech representations in feedforward models. In everyday situations, visual articulatory input would normally be lip information only. However, instrumental methods of transducing tongue motion (e.g., magnetometry, ultrasound, MRI) raise the possibility that visual tongue information may also play a role.

Neurocomputational models of speech production provide a potentially useful framework for understanding the intricacies of AV speech processing. These models seek to provide an integrated explanation for speech processing, incorporated in testable artificial neural networks. Two prominent models include "Directions Into Velocities of Articulators" (DIVA) (Guenther and Perkell, 2004; Guenther, 2006; Guenther et al., 2006; Guenther and Vladusich, 2012) and "ACTion" (ACT) (Kröger et al., 2009). These models assume as input an abstract speech sound unit (a phoneme, syllable, or word), and generate as output both articulatory and auditory representations of speech. The systems operate by computing neural layers (or "maps") as distributed activation patterns. Production of an utterance involves fine-tuning between speech sound maps, sensory maps, and motor maps, guided by feedforward (predictive) processes and concurrent feedback from the periphery. Learning in these models critically relies on forward and inverse processes, with the internal speech model iteratively strengthened by the interaction of feedback information.

Researchers have used neurocomputational frameworks to gain important insights about speech and language disorders, including apraxia of speech (AOS) in adults (Jacks, 2008; Maas et al., 2015), childhood apraxia (Terband et al., 2009; Terband and Maassen, 2010), developmental speech sound disorders (Terband et al., 2014a,b), and stuttering (Max et al., 2004; Civier et al., 2010). For example, DIVA simulations have been used to test the claim that apraxic disorders result from relatively preserved feedback (and impaired feed-forward) speech motor processes (Civier et al., 2010; see also Maas et al., 2015). These neurocomputational modeling-based findings correspond with largely positive results from visual augmented feedback intervention studies for individuals with AOS (see Katz and McNeil, 2010 for review; also, Preston and Leaman, 2014). Overall, these intervention findings have suggested that visual augmented feedback of tongue movement can help remediate speech errors in individuals with AOS, presumably by strengthening the internal model. Other clinical studies have reported that visual feedback can positively influence the speech of individuals with a variety of speech and language problems in children and adults, including articulation/phonological disorders, residual sound errors, and dysarthria. This research has included training with electropalatography (EPG) (Hardcastle et al., 1991; Dagenais, 1995; Goozee et al., 1999; Hartelius et al., 2005; Nordberg et al., 2011), ultrasound (Bernhardt et al., 2005; Preston et al., 2014) and strain gauge transducer systems (Shirahige et al., 2012; Yano et al., 2015).

Visual feedback training has also been used to study information processing during second language (L2) learning. For example, Levitt and Katz (2008) examined augmented visual feedback in the production of a non-native consonant sound. Two groups of adult monolingual American English speakers were trained to produce the Japanese post-alveolar flap /ɽ/. One group received traditional second language instruction alone and the other group received traditional second language instruction plus visual feedback for tongue movement provided by a 2D EMA system (Carstens AG100, Carstens Medizinelektronik GmbH, Bovenden, Germany, www.articulograph.de). The data were perceptually rated by monolingual Japanese native listeners and were also analyzed acoustically for flap consonant duration. The results indicated improved acquisition and maintenance by the participants who received traditional instruction plus EMA training. These findings suggest that visual information regarding consonant place of articulation can assist second language learners with accent reduction.

In another recent study, Suemitsu et al. (2013) tested a 2D EMA-based articulatory feedback approach to facilitate production of an unfamiliar English vowel (/æ/) by five native speakers of Japanese. Learner-specific vowel positions were computed for each participant and provided as feedback in the form of a multiple-sensor, mid-sagittal display. Acoustic analysis of subjects' productions indicated that acoustic and articulatory training resulted in significantly improved /æ/ productions. The results suggest feasibility and applicability to vowel production, although additional research will be needed to determine the separable roles of acoustic and articulatory feedback in this version of EMA training.

Recent research has shown that 3D articulography systems afford several advantages over 2D systems: recording in x/y/z dimensions (and two angles), increased accuracy, and the ability to track movement from multiple articulators placed at positions other than tongue midline (Berry, 2011; Kroos, 2012; Stella et al., 2013). As such, visual augmented feedback provided by these systems may offer new insights on information processing during speech production. A preliminary test of a 3D EMA-based articulatory feedback system was conducted by Katz et al. (2014). Monolingual English speakers were asked to produce several series of four CV syllables. Each series contained four different places of articulation, one of which was an alveolar (e.g., bilabial, velar, alveolar, palatal; such as /pa/-/ka/-/ta/-/ja/). A 1-cm target sphere was placed at each participant's alveolar region. Four of the five participants attempted the series with no visible feedback. The fifth subject was given articulatory visual feedback of their tongue movement and requested to "hit the target" during their series production. The results showed that subjects in the no-feedback condition ranged between 50 and 80% accuracy, while the subject given feedback showed 90% accuracy. These preliminary findings suggested that the 3D EMA system could successfully track lingual movement for consonant feedback purposes, and that feedback could be used by talkers to improve consonantal place of articulation during speech.

A more stringent test of whether 3D visual feedback can modify speech production would involve examining how individuals perform when they must achieve an unfamiliar articulatory target, such as a foreign speech sound. Therefore, in the present experiment we investigated the accuracy with which healthy monolingual talkers could produce a novel, non-English, speech sound (articulated by placing the tongue blade at the palatal region of the oral cavity) and whether this gesture could benefit from short-term articulatory training with visual feedback.

MATERIALS AND METHODS

This study was conducted in accordance with the Department of Health and Human Services regulations for the protection of human research subjects, with written informed consent received from all subjects prior to the experiment. The protocol for this research was approved by the Institutional Review Board at the University of Texas at Dallas. Consent was obtained from all subjects appearing in audio, video, or figure content included in this article.

Participants and Stimuli

Five college-age subjects (three male, two female) with General American English (GAE) accents participated in this study. All talkers were native speakers of English with no speech, hearing, or language disorders. Three participants had elementary speaking proficiency with a foreign language (M03, F02: *Spanish*; F01: *French*). Participants were trained to produce a novel consonant in the /ɑCɑ/ context while an electromagnetic articulograph system recorded lingual movement. For this task, we selected a speech sound not attested as a phoneme among the world's languages: a voiced, coronal, palatal stop. Unlike palatal stops produced with the tongue body, found in languages such as Czech (/c/ and /ɟ/), subjects were asked to produce a closure with the tongue anterior (tip/blade) contacting the hard palate. This sound is similar to a voiced retroflex alveolar /ɖ/, but is articulated in the palatal, not immediately post-alveolar region. As such, it may be represented in the IPA as a backed, voiced retroflex stop: /ɖ̠/. Attested cases appear rarely in the world's languages and only as allophones. For instance, Dart (1991) notes some speakers of O'odham (Papago) produce voiced palatal sounds with (coronal) laminal articulation, instead of the more usual tongue body articulation (see Supplementary Materials for a sample sound file used in the present experiment).

Stimuli were elicited in blocks of 10 /ɑCɑ/ production attempts under a single-subject ABA design. Initially, the experimental protocol called for three pre-training, three training, and three post-training blocks from each subject (for a total of 90 productions). However, because data for this study were collected as part of a larger investigation of stop consonant productions, there was some subject attrition and reduced participation for the current experiment. Thus, the criterion for completion of the experiment was changed to a minimum of one block of baseline (no feedback) probes, 2–3 blocks of visual feedback training, and 1–3 blocks of post-feedback probes, for a total of 40–80 productions from each participant. All trials were conducted within a single experimental session lasting approximately 15 min.

Procedure

Training sessions were conducted in a quiet testing room at the University of Texas at Dallas. Each participant was seated next to the Wave system, facing a computer monitor located approximately 1 m away. Five sensors were glued to the subject's tongue using a biocompatible adhesive: one each at tongue tip (∼1 cm posterior to the apex), tongue middle (∼3 cm posterior to apex), tongue back (∼4 cm posterior to the apex), and both left and right tongue lateral positions. Sensors were also attached to a pair of glasses worn by the subject to establish a frame of reference for head movement. A single sensor was taped on the center of the chin to track jaw movement.

Visual Feedback Apparatus

External visual feedback for lingual movement was provided to subjects using a 3D EMA-based system (*Opti-Speech*, Vulintus LLC, Sachse, Texas, United States, http://www.vulintus.com/). This system works by tracking speech movement with a magnetometer (*Wave*, Northern Digital Incorporated, Waterloo, Ontario, Canada). An interface allows users to view their current tongue position (represented by an image consisting of flesh-point markers and a modeled tongue surface) within a transparent head with a moving jaw. Small blue spheres mark different regions on the animated tongue (tongue tip, tongue middle, tongue back, or tongue left/right lateral). Users may adjust the visibility of these individual markers and/or select or deselect "active" markers for speech training purposes. Articulatory targets, shown on the screen as semi-transparent red or orange spheres, can be placed by the user in the virtual oral cavity. The targets change color to green when the active marker enters, indicating correct tongue position, thus providing immediate visual feedback for place of articulation (see Katz et al., 2014 for more information). The target size and "hold time on target" can be varied by the user to make the target matching task easier or harder. An illustration of the system is shown in **Figure 1**.

Pronunciation Training

The backed palatal stop consonant /ɖ̠/ is produced by making a closure between the tongue tip and hard palate. Therefore, the tongue tip marker was designated as the active marker for this study. A single target was placed at the palatal place of articulation to indicate where the point of maximum constriction should occur during the production of /ɖ̠/. To help set the target, participants were requested to press their tongue to the roof of their mouth, allowing the tongue sensors to conform to the contours of the palate. The experimenter then placed the virtual target at the location of the tongue middle sensor, which was estimated to correspond to the palatal (typically, pre-palatal) region. Based on previous work (Katz et al., 2014), we

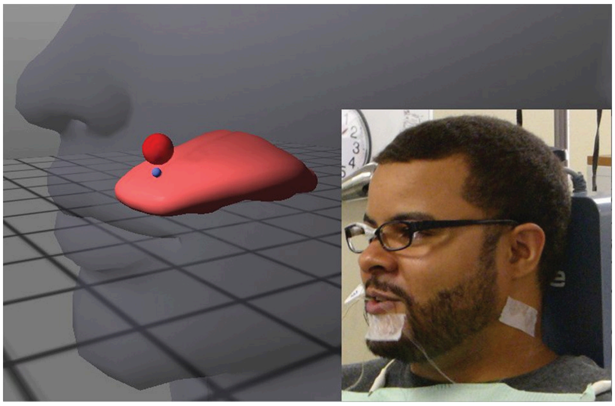

FIGURE 1 | Illustration of the *Opti-Speech* system, with subject wearing sensors and head-orientation glasses (lower right insert). A sample target sphere, placed in this example at the subject's alveolar ridge, is shown in red. A blue marker indicates the tongue tip/blade (TT) sensor.

selected a target sphere of 1.00 cm in volume, with no hold time.

The current experiment was conducted as part of a larger study investigating stop consonant production that employed visual feedback for training purposes. As such, by the start of the experiment each participant had received an opportunity to accommodate to the presence of the Wave sensors on the tongue and to practice speaking English syllables and words under visual feedback conditions for approximately 25–30 min. In order to keep practice conditions uniform in the actual experiment, none of these warmup tasks involved producing a novel, non-English sound.

For the present experiment, participants were trained to produce the voiced, coronal, palatal stop, /ɖ̠/. The investigator (SM) described the sound to subjects as "sound[ing] like a '*d*,' but produced further back in the mouth." A more precise articulatory explanation was also provided, instructing participants to feel along the top of their mouth from front to back to help identify the alveolar ridge. Participants were then told to "place the tip of [their] tongue behind the alveolar ridge and slide it backwards to meet with the roof, or palate, of the mouth." The investigator, a graduate student with a background in phonetics instruction, produced three repetitions of /ɑɖ̠ɑ/ (live) for participants to imitate. Each participant was allowed to practice making the novel consonantal sound 3–5 times before beginning the no-feedback trial sessions. This practice schedule was devised based on pilot data suggesting 3–5 practice attempts were sufficient for participants to combine the articulatory, modeled, and feedback information to produce a series of successive "best attempts" at the novel sound. Throughout the training procedure, the investigator provided generally encouraging comments. In addition, if an attempt was judged perceptually to be off-target (e.g., closer to an English /d/ or the palatalized alveolar stop, /d^j/), the investigator pointed out the error and repeated the (articulatory) instructions.

When the participant indicated that he/she understood all of the instructions, pre-training (baseline) trials began. After each block of attempts, participants were given general feedback about their performance and the instructions were reiterated if necessary. Once all pre-training sessions were completed, the participant was informed that the *Opti-Speech* visual feedback system would now be used to help them track their tongue movement. Subjects were instructed to use the tongue model as a guide for producing the palatal sound by moving the tongue tip upwards and backwards until the tongue tip marker entered the palatal region and the target lit up green, indicating success (see **Figure 2**). Each participant was allowed three practice attempts at producing the novel consonant while simultaneously watching the tongue model and aiming for the virtual target.

After completing the training sessions, the subject was asked to once again attempt to produce the sound with the visual feedback removed. No practice attempts were allowed between the training and post-training trial sessions. During all trials, the system recorded the talker's kinematic data, including a record of target hits (i.e., accuracy of the tongue-tip sensor entering the subject's palatal zone). The experiments were also audio- and video-recorded.

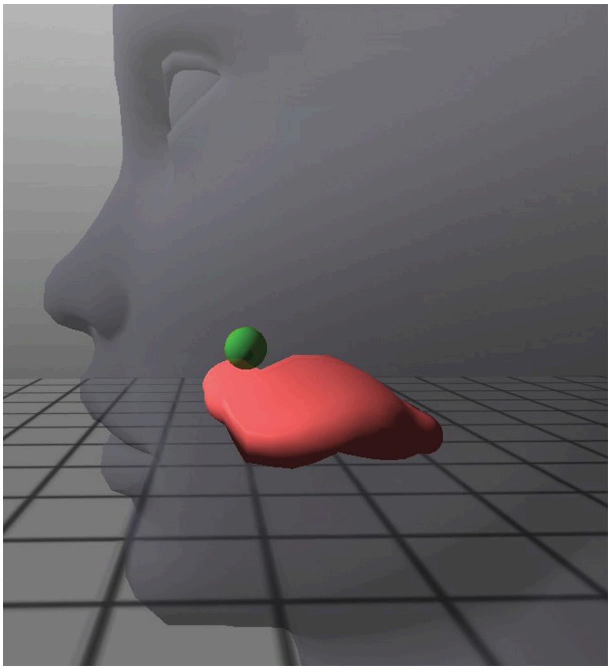

FIGURE 2 | Close-up of tongue avatar during a "hit" for the production of the voiced, retroflex, palatal stop consonant. The target sphere lights up green, providing visual feedback for the correct place of articulation.

RESULTS

Kinematic Results

All participants completed the speaking task without noticeable difficulty. Speakers' accuracy in achieving the correct articulation was measured as the number of hit targets out of the number of attempts in each block. Talker performance is summarized in **Figure 3**, which shows accuracy at the baseline (pre-training), visual feedback (shaded), and post-feedback (post-training) probes.

All talkers performed relatively poorly at baseline phase, ranging from 0 to 50% ($x = 12.6\%$, $sd = 14.1\%$) accuracy. Each participant showed a rapid increase in accuracy during the visual feedback phase (shaded), ranging from 50 to 100% ($x = 74.9\%$, $sd = 15.6$). These gains appeared to be maintained during the post-feedback probes, with scores ranging from 70 to 100% ($x = 85.3\%$, $sd = 12.8\%$). Group patterns were examined using two-way paired *t*-tests. The results indicated a significant difference between pre-training and training phases, $t_{(4)} = 8.73$, $p < 0.001$, and pre-training and post-training phases, $t_{(4)} = 14.0$, $p < 0.001$. No significant difference was found between training and post-training, $t_{(4)} = 1.66$, *ns*. This pattern suggests acquisition during the training phase, and maintenance of learned behavior immediately post-training.

An effect size for each subject was computed using the Percentage of Non-overlapping Data (PND) method described by Scruggs et al. (1987). This non-parametric analysis compares points of non-overlap between baseline and successive

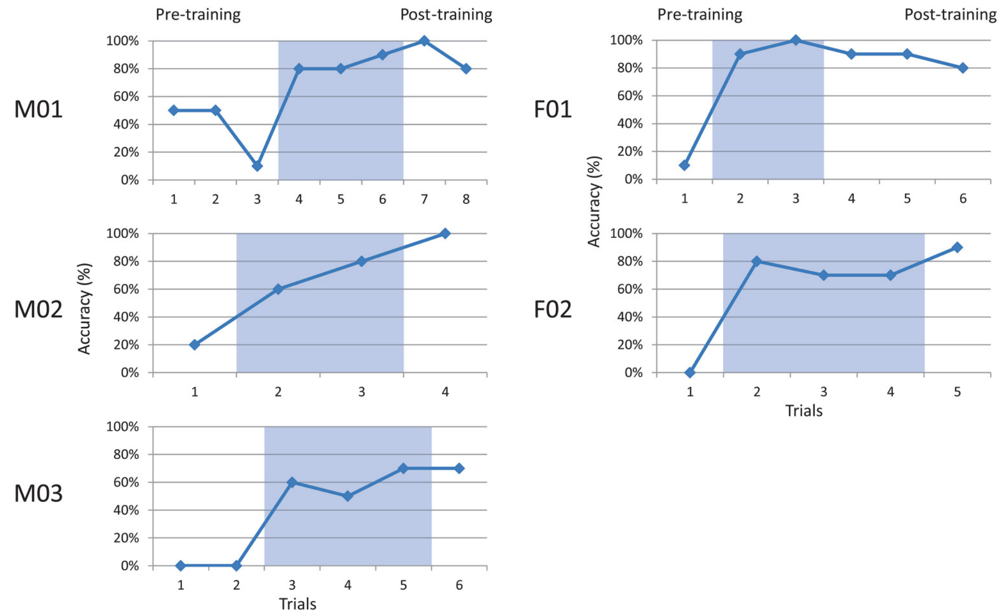

FIGURE 3 | Accuracy for five talkers producing a coronal palatal stop. Shaded regions indicate visual feedback conditions. Baseline (pre-training) and post-training phases are also indicated.

intervention phases, and criteria are suggested for interpretation (Scruggs et al., 1986). Using this metric, all of the subjects' patterns were found to be greater than 90% (*highly effective*) for comparisons of both pre-training vs. training, and pre-training vs. post-training.

Acoustic Results

In order to corroborate training effects, we sought acoustic evidence of coronal (tongue blade) palatal stop integrity. This second analysis investigated whether the observed improvement in talkers' articulatory precision resulting from training would be reflected in patterns of the consonant burst spectra. Short-term spectral analyses were obtained at the moment of burst release (Stevens and Blumstein, 1975, 1978). Although, burst spectra may vary considerably from speaker to speaker, certain general patterns may be noted. Coronals generally have energy distribution across the whole spectrum, with at least two peaks between 1.2 and 3.6 kHz), termed "diffuse" in the feature system of Jakobson et al. (1952). Also, coronals typically result in relatively higher-frequency spectral components than articulations produced by lips or the tongue body, and these spectra are therefore described as being "acute" (Jakobson et al., 1952; Hamann, 2003) or "diffuse-rising" (Stevens and Blumstein, 1978).

Burst frequencies vary as a function of the length of the vocal tract anterior to the constriction. Thus, alveolar constriction results in a relatively high burst, ranging from approximately 2.5 to 4.5 kHz (e.g., Reetz and Jongman, 2009), while velar stops, having a longer vocal tract anterior to the constriction, produce lower burst frequencies (ranging from approximately 1.5 to 2.5 kHz). Since palatal stops are produced with a constriction located between the alveolar and velar regions, palatal stop bursts may be expected to have regions of spectral prominence between the two ranges, in the 3.0–5.0 kHz span. Acoustic analyses of Czech or Hungarian velar and palatal stops generally support this view. For instance, Keating and Lahiri (1993) note that the Hungarian palatal stop /ca/ spectrum slopes up to its highest peak "at 3.0–4.0 kHz or ever higher," but otherwise show "a few peaks of similar amplitude which together dominate the spectrum in a single broad region" (p. 97). A study by Dart (1991) obtained palatographic and spectral data for O'odham (Papago) voiced palatal sounds produced with laminal articulation. Analysis of the burst spectra for these (O'odham) productions revealed mostly diffuse rising spectra, with some talkers showing "a high amplitude peak around 3.0–5.0 Hz" (p. 142).

For the present experiment, three predictions were made: (1) palatal stop consonant bursts prior to training will have diffuse rising spectra with characteristic peaks in the 3.0–5.0 kHz range, and (2) following training, these spectral peaks will shift downwards, reflecting a more posterior constriction (e.g., from an alveolar toward a palatal place of articulation), and (3) post-training token-to-token variability should be lower than at baseline, reflecting increased articulatory ability.

Spectral Analysis

Talkers' consonantal productions were digitized and analyzed using PRAAT (Boersma and Weenink, 2001) with a scripting procedure using linear predictive coding (LPC) analysis. A cursor was placed at the beginning of the consonant burst of each syllable and a 12 ms Kaiser window was centered over the stop transient. Autocorrelation-based LPC (24 pole model, +6 dB pre-emphasis) yielded spectral sections. Overlapping plots of subjects' repeat utterances were obtained for visual inspection, with spectral peaks recorded for analysis.

Figure 4 shows overlapping plots of spectra obtained pre- and post-EMA training for 4/5 talkers. Plots containing (RMS) averages for pre-training (incorrect) and post-training (correct) spectra are also shown, for comparison. Spectra for talker F01 could not be compared because this talker's initial productions were realized as CV syllables (instead of VCV), and differing vowel context is known to greatly affect burst consonant spectral characteristics (Stevens, 2008).

Results revealed mixed support for the experimental predictions. Similar to previous reports (e.g., Dart, 1991), there were considerable differences in the shapes of the burst spectral patterns from talker to talker. Three of the four talkers' spectra (M01, M02, and M03) were diffuse, having at least two peaks between 1.2 and 3.6 kHz, while the spectra of talker F01 had peaks in a mid-frequency ("compact") range of 2.0–3.0 kHz. Patterns of spectral tilt for all speakers were generally falling (instead of rising, as expected).

The prediction that 3.0–5.0 kHz spectral peak frequencies would lower following training was not uniformly obtained. Because standard deviations were relatively high and there was much inter-talker variability, the data are summarized, rather than tested statistically.

Talker M01's data had six peaks pre-treatment ($x = 3967$; $sd = 596$) and five peaks post-training ($x = 4575$; $sd = 281$). Talker M02's productions yielded five peaks pre-training ($x = 3846$; $sd = 473$) and nine peaks post-training ($x = 3620$; $sd = 265$). Talker M03 had six peaks pre-training ($x = 4495$ $sd = 353$) and nine peaks post-training ($x = 3687$; $sd = 226$). The spectra of talker F02 had peaks in a mid-frequency ("compact") range of approximately 2.0–3.0 kHz. This talker's spectral peak values did not shift with training (pre-training: $x = 2359$ Hz, $sd = 139$ Hz; post-training: $x = 2390$ Hz, $sd = 194$ Hz). In summary, talkers M03 and M02 showed the expected pattern of spectra peak

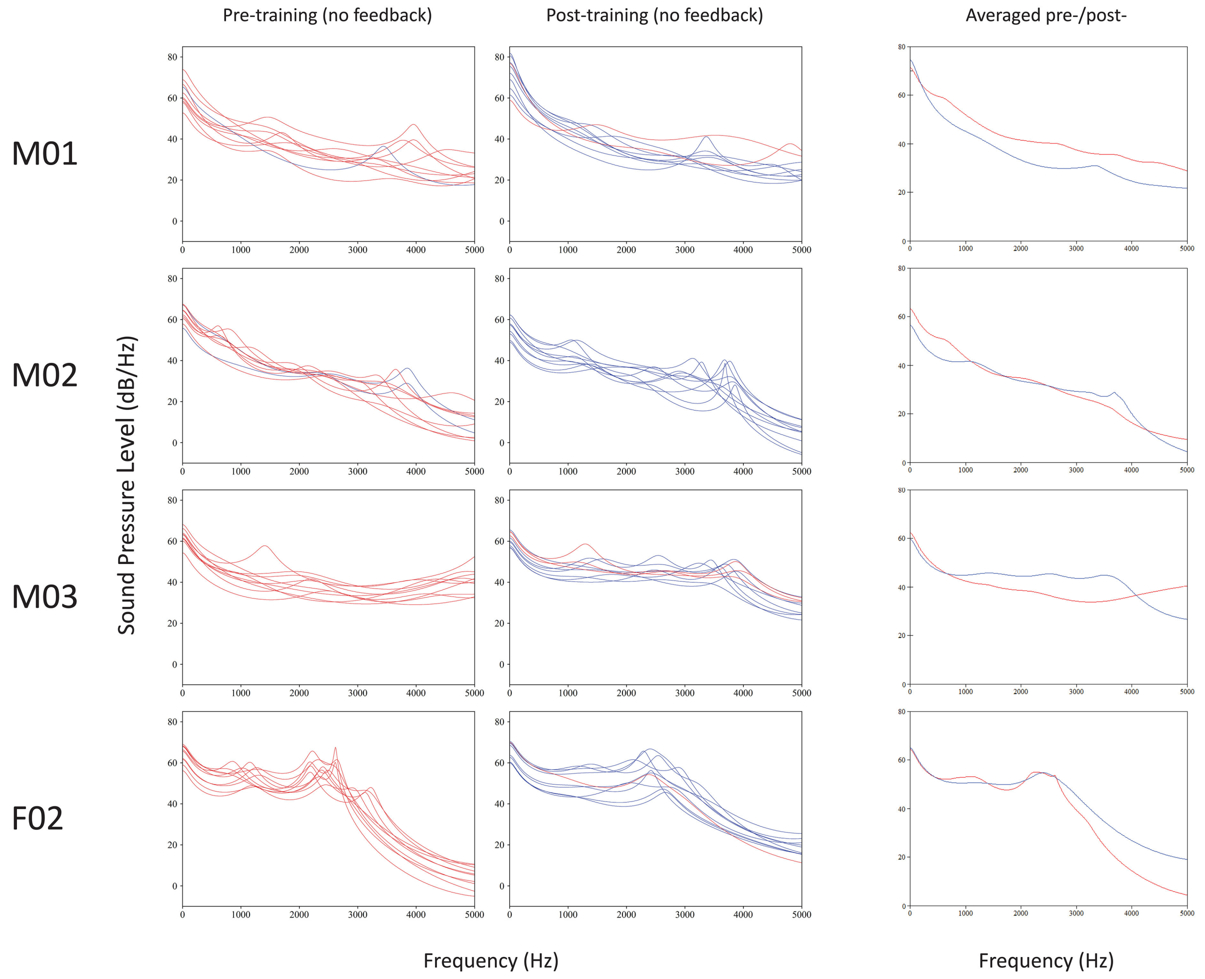

FIGURE 4 | Overlapping plots of short-term spectra for bursts of voiced, coronal, palatal stops produced before and after EMA training. Correct place of articulation (hits) are marked in blue, and errors (misses) in red. Computed averages of incorrect pre-training (red) and correct post-training (blue) spectra are shown at right, for comparison.

lowering, F02 showed no training-dependent changes, and M01 showed a pattern in the opposite direction.

Of the talkers with spectra data available, three (M01, M02, and M03) showed marked reduction in variability (i.e., reduced standard deviation values) from pre-training to post-training, suggesting that training corresponded with increased production consistency. However, this was not the case for talker F02, whose mid-range spectral peaks showed a slight increase in variability after training.

DISCUSSION

Five English-speaking subjects learned a novel consonant (a voiced, coronal, and palatal stop) following a brief training technique involving visual augmented feedback of tongue movement. The results of kinematic analyses indicate that real-time visual (articulatory) feedback resulted in improved accuracy of consonant place of articulation. Articulatory feedback training for place of articulation corresponded with a rapid increase in the accuracy of tongue tip spatial positioning, and post-training probes indicated (short-term) retention of learned skills.

Acoustic data for talkers' burst spectra obtained pre- and post-training only partially confirmed the kinematic findings, and there were a number of differences noted from predictions. First, for those talkers that showed diffuse spectra (e.g., with two peaks between 1.2 and 3.6 kHz), the spectra were falling, instead of rising. This may have been due to a number of possible factors, including the current choice of a Kaiser window for spectral analysis. Some of the original studies, such as those which first noted the classic "diffuse rising" patterns in spectral slices, fitted half-Hamming windows over the burst to obtain optimum pre-emphasis for LPC analysis (e.g., Stevens and Blumstein, 1978). Second, talker F02 showed mid-range ("compact") spectral peaks ranging between 2.0 and 3.0 kHz. This may be due to tongue shape, which can affect the affect spectral characteristics of the stop burst. For example, laminal (tongue blade) articulation results in relatively even spectral spread, while apical (tongue-tip) articulation results in strong mid-frequency peaks (Ladefoged and Maddieson, 1996) and less spread (Fant, 1973). In the present data, the spectra of talker F02 fits that pattern of a more apical production.

Despite individual differences, there was some evidence supporting the notion of training effects in the acoustic data. Chiefly, the three subjects with diffuse spectra (M01, M02, and M03) showed decreased variability (lowered standard deviations) following training, suggesting stabilized articulatory behavior. Although the current data are few, they suggest that burst spectra variability may be a useful metric to be explored in future studies.

It was predicted that spectral peaks in the 3.0–5.0 kHz range would lower in frequency as talkers improved their place of articulation, with training. However, the findings do not generally support this prediction: Talker M03 showed this pattern, M02 showed a trend, F02 showed no differences, and M01 trended in the opposite direction, with higher spectral peaks after training. Since the kinematic data establish that all talkers significantly increased tongue placement accuracy post-training, we speculate that several factors affecting burst spectra (e.g., tongue shape, background noise, or room acoustics) may have obscured any such underlying spectral shifts for the talkers. Future research should examine how burst spectra may be best used to evaluate outcomes in speech training studies.

The current kinematic data replicate and extend the findings of Ouni (2013) who found that talkers produced tongue body gestures more accurately after being exposed to a short training session of real-time ultrasound feedback (post-test) than when recorded at baseline (pre-test). The present results are also consistent with earlier work from our laboratory which found that monolingual English speakers showed faster and more effective learning of the Japanese post-alveolar flap, /ɽ/ using EMA-based visual feedback, when compared with traditional Japanese pronunciation instruction (Levitt and Katz, 2008). Taken together with the experimental data from this study, there is evidence that EMA-provided articulatory visual feedback may provide a means for helping L2 learners improve novel consonant distinctions.

However, a number of caveats must be considered. First, the current data are limited and the study should therefore be considered preliminary. The number of subjects tested was few ($n = 5$). Also, since the consonant trained, /ɖ̠/, is not a phoneme in any of the world's language, it was not possible to include perceptual data, such as native listener judgments (e.g., Levitt and Katz, 2010). Additional data obtained from more talkers will therefore be required before any firm conclusions can be drawn concerning the relation to natural language pronunciation.

Second, real-time (live) examples were given to subjects by the experimenter (SM) during the training phase, allowing for the possibility of experimenter bias. This procedure was adopted to simulate a typical second-language instruction setting, and care was taken to produce consistent examples, so as to not introduce "unfair" variability at the start of the experiment. Nevertheless, in retrospect it would have been optimal to have included a condition in which talkers were trained with pre-recorded examples, to eliminate this potential bias.

Third, since articulatory training is assumed to draw on principles of motor learning, several experimental factors must be controlled before it is possible to conclude that a given intervention is optimal for a skill being acquired, generalized, or maintained (e.g., Maas et al., 2008; Bislick et al., 2012; Schmidt and Lee, 2013; Sigrist et al., 2013). For example, Ballard et al. (2012) conducted a study in which a group of English talkers was taught the Russian trilled /r/ sound using an EPG-based visual feedback system. In a short-term (five session) learning paradigm, subjects practiced in conditions either with continuous visual feedback provided by an EPG system, or were given no visual feedback. The results suggested that providing kinematic feedback continually though treatment corresponded with lower skill retention. This finding suggests that speech training follows the principle that kinematic feedback is most beneficial in the early phases of training, but may interfere with long-term retention if provided throughout training (Swinnen et al., 1993; Hodges and Franks, 2001; Schmidt and Lee, 2013). A pattern in the current data also potentially supports this principle. Three of the five participants (M01, M03, and F02) reached their maximum performance in the post-training phase, immediately after the feedback was removed. While this pattern was not statistically significant, it may suggest some interference effects from the ongoing feedback used. Future research should examine factors such as feedback type and frequency in order to better improve speech sound learning.

The current findings support the notion of a visual feedback pathway during speech processing, as proposed in the ACT neurocomputational model of speech production (Kröger and Kannampuzha, 2008). Similar to the DIVA model, ACT relies on feedforward and feedback pathways between distributed neural activation patterns, or maps. ACT includes explicit provisions for separate visual and auditory information processing. In **Figure 5**, we present a simplified model of ACT (adapted from Kröger et al., 2009) with (optional) modifications added to highlight pathways for external and internal audiovisual input. Since people do not ordinarily rely on visual feedback of tongue movement, these modifications explain how people learn under conditions of augmented feedback, rather than serving as key components of everyday speech.

The external input route (dotted circle on the right) indicates an outside speech source, including speech that is produced while hearing/observing human talkers or a computerized training agent (e.g., BALDI, ARTUR, ATH, or Vivian). The input audio and visual data are received, preprocessed, and relayed as input to respective unimodal maps. These maps yield output to a multimodal phonetic map that also receives (as input) information from a somatosensory map and from a phonemic map. Reciprocal feedback connections between the phonetic map, visual-phonetic processing, and auditory-phonetic processing modules can account for training effects from computerized training avatars. These pathways would presumably also be involved in AV model-learning behavior, including lip-reading abilities (see Bernstein and Liebenthal, 2014 for review) and compensatory tendencies noted in individuals with left-hemisphere brain damage, who appear to benefit from visual entrainment to talking mouths other than their own (Fridriksson et al., 2012).

In the (internal) visual feedback route (dotted arrows), a talker's own speech articulation is observed during production.

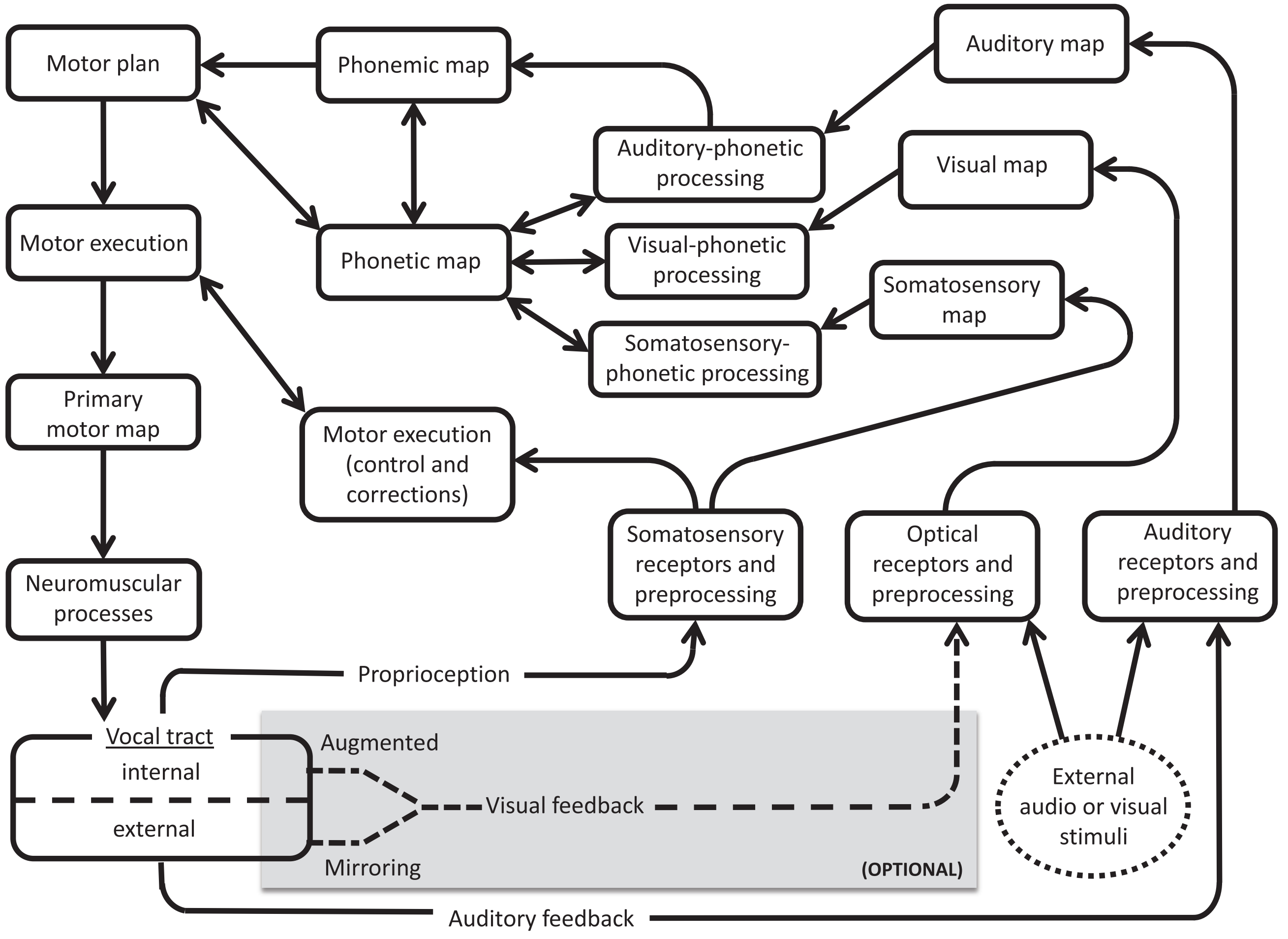
Motor plan
Phonemic map
Auditory map
Auditory-phonetic processing
Visual map
Motor execution
Phonetic map
Visual-phonetic processing
Somatosensory map
Somatosensory-phonetic processing
Primary motor map
Motor execution (control and corrections)
Somatosensory receptors and preprocessing
Optical receptors and preprocessing
Auditory receptors and preprocessing
Neuromuscular processes
Proprioception
Vocal tract
internal
external
Augmented
Visual feedback
Mirroring
(OPTIONAL)
External audio or visual stimuli
Auditory feedback

FIGURE 5 | Simplified version of ACT model (Kröger and Kannampuzha, 2008), showing input pathways for external audiovisual stimuli (oval at bottom right) and optional feedback circuits to the vocal tract (shaded box at bottom). Visual feedback (dotted line) is provided by either external (mirroring) or internal (instrumental augmented) routes.

This may include simple mirroring of the lips and jaw, or instrumentally augmented visualizations of the tongue (via EMA, ultrasound, MRI, or articulatory inversion systems that convert sound signals to visual images of the articulators; e.g., Hueber et al., 2012). The remaining audio and visual preprocessing and mapping stages are similar between this internal route and the external (modeled) pathways. The present findings of improved consonantal place of articulation under conditions of visual (self) feedback training supports this internal route and the role of body sense/motor familiarity. This internal route may also play a role in explaining a number of other phenomena described in the literature, including the fact that talkers can discern between natural and unnatural tongue movements displayed by an avatar (Engwall and Wik, 2009), and that training systems based on a talkers' own speech may be especially beneficial for L2 learners (see Felps et al., 2009 for discussion).

The actual neurophysiological mechanisms underlying AV learning and feedback are currently being investigated. Recent work on oral somatosensory awareness suggests people have a unified "mouth image" that may be qualitatively different from other parts of the body (Haggard and de Boer, 2014). Since visual feedback does not ordinarily play a role in mouth experiences, other attributes, such as self-touch, may play a heightened role. For instance, Engelen et al. (2002) note that subjects can achieve high accuracy in determining the size of ball-bearings placed in the mouth, but show reduced performance when fitted with a plastic palate. This suggests that relative movement of an object between tongue and palate is important in oral size perception. We speculate that visual feedback systems rely in part on oral self-touch mechanism (particularly for consonant production), by visually guiding participants to the correct place of articulation, at which point somatosensory processes take over. This mechanism may prove particularly important for consonants, as opposed to vowels, which are produced with less articulatory contact.

Providing real-time motor feedback may engage different cortical pathways than are recruited in learning systems that employ more traditional methodologies. For example, Farrer et al. (2003) conducted positron emission tomography (PET) experiments in which subjects controlled a virtual hand on a screen under conditions ranging from full control, to partial control, to a condition where another person controlled the hand and there was no control. The results showed right inferior parietal lobule activation when subjects felt least in control of the hand, with reverse covariation in the insula. A crucial aspect here is corporeal identity, the feeling of one' own body, in order to determine motor behavior in the environment. Data suggest that body awareness is supported by a large network of neurological structures including parietal and insular cortex, with primary and secondary somatosensory cortex, insula, and posterior parietal cortex playing specific roles (see Daprati et al., 2010 for review). A region of particular interest is the right inferior parietal lobule (IPL), often associated to own-body perception and other body discrimination (Berlucchi and Aglioti, 1997; Farrer et al., 2003; Uddin et al., 2006). Additional neural structures that likely play a role in augmented feedback training systems include those associated with reward dependence during behavioral performance, including lateral prefrontal cortex (Pochon et al., 2002; Liu et al., 2011; Dayan et al., 2014). As behavioral data accrue with respect to both external (mirroring) and internal ("tongue reading") visual speech feedback, it will be important to also describe the relevant neural control structures, in order to best develop more complete models of speech production.

In summary, we have presented small-scale but promising results from an EMA-based feedback investigation suggesting that augmented visual information concerning one's own tongue movements boosts skill acquisition during the learning of consonant place of articulation. Taken together with other recent data (e.g., Levitt and Katz, 2010; Ouni, 2013; Suemitsu et al., 2013) the results may have potentially important implications for models of speech production. Specifically, distinct AV learning mechanisms (and likely, underlying neural substrates) appear to be engaged for different types of CAPT systems, with interactive, on-line, eye-to-tongue coordination involved in systems such as *Opti-Speech* (and perhaps *Vizart3D*, Hueber et al., 2012) being arguably different than processing involved in using external avatar trainers, such as ARTUR, BALDI, ATH, or Vivian. These different processing routes may be important when interpreting other data, such as the results of real-time, discordant, cross-modal feedback (e.g., McGurk effect). Future, studies should focus on extending the range of speech sounds, features, and articulatory structures trained with real-time feedback, with a focus on vowels as well as consonants (see Mehta and Katz, 2015). As findings are strengthened with designs that systematically test motor training principles, the results may open new avenues for understanding how AV information is used in speech processing.

AUTHOR CONTRIBUTIONS

WK and SM designed the experiments. SM recruited the participants and collected the data. WK and SM performed the kinematic analysis. WK conducted the spectral analysis. WK and SM wrote the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the University of Texas at Dallas Office of Sponsored Projects, the UTD Callier Center Excellence in Education Fund, and a grant awarded by NIH/NIDCD (R43 DC013467). We thank the participants for volunteering their time and Carstens Medezinelektronik GmbH for material support toward our research. We would also like to thank Marcus Jones, Amy Berglund, Cameron Watkins, Bill Watts, and Holle Carey for their contributions in apparatus design, data collection, data processing, and other support for this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnhum.2015.00612

REFERENCES

Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci.* 28, 105–124. doi: 10.1017/S0140525X05000038

Arnold, P., and Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.* 92(Pt 2), 339–355. doi: 10.1348/000712601162220

Badin, P., Elisei, F., Bailly, G., and Tarabalka, Y. (2008). "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data," in *Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098)*, eds F. J. Perales and R. B. Fisher, (Berlin; Heidelberg: Springer Verlag), 132–143. doi: 10.1007/978-3-540-70517-8_14

Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Commun.* 52, 493–503. doi: 10.1016/j.specom.2010.03.002

Ballard, K. J., Smith, H. D., Paramatmuni, D., McCabe, P., Theodoros, D. G., and Murdoch, B. E. (2012). Amount of kinematic feedback affects learning of speech motor skills. *Motor Contr.* 16, 106–119.

Berlucchi, G., and Aglioti, S. (1997). The body in the brain: neural bases of corporeal awareness. *Trends Neurosci.* 20, 560–564. doi: 10.1016/S0166-2236(97)01136-3

Bernhardt, B., Gick, B., Bacsfalvi, P., and Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clin. Linguist. Phon.* 19, 605–617. doi: 10.1080/02699200500114028

Bernstein, L. E., and Liebenthal, E. (2014). Neural pathways for visual speech perception. *Front. Neurosci.* 8:386. doi: 10.3389/fnins.2014.00386

Berry, J. J. (2011). Accuracy of the NDI wave speech research system. *J. Speech Lang. Hear. Res.* 54, 1295–1301. doi: 10.1044/1092-4388(2011/10-0226)

Bislick, L. P., Weir, P. C., Spencer, K., Kendall, D., and Yorkston, K. M. (2012). Do principles of motor learning enhance retention and transfer of speech skills? A systematic review. *Aphasiology* 26, 709–728. doi: 10.1080/02687038.2012.676888

Boersma, P., and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

Civier, O., Tasko, S. M., and Guenther, F. H. (2010). Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *J. Fluency Disord.* 35, 246–279. doi: 10.1016/j.jfludis.2010.05.002

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., and Hari, R. (2000). Speaking modifies voice−evoked activity in the human auditory cortex. *Hum. Brain Mapp.* 9, 183–191. doi: 10.1002/(SICI)1097-0193(200004)9:4<183::AID-HBM1>3.0.CO;2-Z

D'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J. J., and Fadiga, L. (2014). Vision of tongue movements bias auditory speech perception. *Neuropsychologia* 63, 85–91. doi: 10.1016/j.neuropsychologia.2014.08.018

Dagenais, P. A. (1995). Electropalatography in the treatment of articulation/phonological disorders. *J. Commun. Disord.* 28, 303–329. doi: 10.1016/0021-9924(95)00059-1

Daprati, E., Sirigu, A., and Nico, D. (2010). Body and movement: consciousness in the parietal lobes. *Neuropsychologia* 48, 756–762. doi: 10.1016/j.neuropsychologia.2009.10.008

Dart, S. N. (1991). *Articulatory and Acoustic Properties of Apical and Laminal Articulations, Vol. 79.* Los Angeles, CA: UCLA Phonetics Laboratory.

Dayan, E., Hamann, J. M., Averbeck, B. B., and Cohen, L. G. (2014). Brain structural substrates of reward dependence during behavioral performance. *J. Neurosci.* 34, 16433–16441. doi: 10.1523/JNEUROSCI.3141-14.2014

Engelen, L., Prinz, J. F., and Bosman, F. (2002). The influence of density and material on oral perception of ball size with and without palatal coverage. *Arch. Oral Biol.* 47, 197–201. doi: 10.1016/S0003-9969(01)00106-6

Engwall, O. (2008). "Can audio-visual instructions help learners improve their articulation?-an ultrasound study of short term changes," in *Interspeech* (Brisbane), 2631–2634.

Engwall, O., and Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Comput. Assist. Lang. Learn.* 20, 235–262. doi: 10.1080/09588220701489507

Engwall, O., Bälter, O., Öster, A.-M., and Kjellström, H. (2006). "Feedback management in the pronunciation training system ARTUR," in *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (Montreal: ACM), 231–234.

Engwall, O., and Wik, P. (2009). "Can you tell if tongue movements are real or synthesized?" in *Proceedings of Auditory-Visual Speech Processing* (Norwich: University of East Anglia), 96–101.

Erber, N. P. (1975). Auditory-visual perception of speech. *J. Speech Hear. Disord.* 40, 481–492. doi: 10.1044/jshd.4004.481

Fagel, S., and Madany, K. (2008). "A 3-D virtual head as a tool for speech therapy for children," in *Proceedings of Interspeech 2008* (Brisbane, QLD*)*, 2643–2646.

Fant, G. (1973). *Speech Sounds and Features.* Cambridge, MA: The MIT Press.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., and Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *Neuroimage* 18, 324–333. doi: 10.1016/S1053-8119(02)00041-1

Felps, D., Bortfeld, H., and Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Commun.* 51, 920–932. doi: 10.1016/j.specom.2008.11.004

Fridriksson, J., Hubbard, H. I., Hudspeth, S. G., Holland, A. L., Bonilha, L., Fromm, D., et al. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Brain* 135, 3815–3829. doi: 10.1093/brain/aws301

Gentilucci, M., and Corballis, M. C. (2006). From manual gesture to speech: a gradual transition. *Neurosci. Biobehav. Rev.* 30, 949–960. doi: 10.1016/j.neubiorev.2006.02.004

Goozee, J. V., Murdoch, B. E., and Theodoros, D. G. (1999). Electropalatographic assessment of articulatory timing characteristics in dysarthria following traumatic brain injury. *J. Med. Speech Lang. Pathol.* 7, 209–222.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013

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, 280–301. doi: 10.1016/j.bandl.2005.06.001

Guenther, F. H., and Perkell, J. S. (2004). "A neural model of speech production and its application to studies of the role of auditory feedback in speech," in *Speech Motor Control in Normal and Disordered Speech*, eds B. Maassen, R. Kent, H. F. M. Peters, P. Van Lieshout, and W. Hulstijn (Oxford University Press), 29–50.

Guenther, F. H., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguistics* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Gunji, A., Hoshiyama, M., and Kakigi, R. (2001). Auditory response following vocalization: a magnetoencephalographic study. *Clin. Neurophysiol.* 112, 514–520. doi: 10.1016/S1388-2457(01)00462-X

Haggard, P., and de Boer, L. (2014). Oral somatosensory awareness. *Neurosci. Biobehav. Rev.* 47, 469–484. doi: 10.1016/j.neubiorev.2014.09.015

Hamann, S. (2003). *The Phonetics and Phonology of Retroflexes*. Ph.D. dissertation, Netherlands Graduate School of Linguistics, University of Utrecht, LOT, Utrecht.

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. *Int. J. Lang. Commun. Disord.* 26, 41–74. doi: 10.3109/13682829109011992

Hartelius, L., Theodoros, D., and Murdoch, B. (2005). Use of electropalatography in the treatment of disordered articulation following traumatic brain injury: a case study. *J. Med. Speech Lang. Pathol.* 13, 189–204.

Heinks-Maldonado, T. H., Nagarajan, S. S., and Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport* 17, 1375. doi: 10.1097/01.wnr.0000233102.43526.e9

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *J. Cogn. Neurosci.* 21, 1229–1243. doi: 10.1162/jocn.2009.21189

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. *Lang. Cogn. Processes* 25, 749–776. doi: 10.1080/01690961003595572

Hodges, N. J., and Franks, I. M. (2001). Learning a coordination skill: interactive effects of instruction and feedback. *Res. Q. Exerc. Sport* 72, 132–142. doi: 10.1080/02701367.2001.10608943

Houde, J. F., Nagarajan, S. S., Sekihara, K., and Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *J. Cogn. Neurosci.* 14, 1125–1138. doi: 10.1162/089892902760807140

Hueber, T., Ben-Youssef, A., Badin, P., Bailly, G., and Elisei, F. (2012). "Vizart3D: retour articulatoire visuel pour l'aide à la pronunciation," in *29e Journées d'Études sur la Parole (JEP-TALN-RECITAL'2012), Vol.* 5, 17–18.

Jacks, A. (2008). Bite block vowel production in apraxia of speech. *J. Speech Lang. Hear. Res.* 51, 898–913. doi: 10.1044/1092-4388(2008/066)

Jakobson, R., Fant, G., Halle, M., Jakobson, R. J., Fant, R. G., and Halle, M. (1952). *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates.* Technical Report, Acoustics Laboratory. No. 13. MIT.

Katz, W. F., Campbell, T. F., Wang, J., Farrar, E., Eubanks, J. C., Balasubramanian, A., et al. (2014). "Opti-speech: A real-time, 3D visual feedback system for speech training," in *Procceedings of Interspeech*. Available online at: https://www.utdallas.edu/~wangjun/paper/Interspeech14_opti-speech.pdf

Katz, W. F., and McNeil, M. R. (2010). Studies of articulatory feedback treatment for apraxia of speech based on electromagnetic articulography. *SIG 2 Perspect. Neurophysiol. Neurogenic Speech Lang. Disord.* 20, 73–79. doi: 10.1044/nnsld20.3.73

Keating, P., and Lahiri, A. (1993). Fronted velars, palatalized velars, and palatals. *Phonetica* 50, 73–101. doi: 10.1159/000261928

Kohler, E., Keysers, C., Umiltá, M. A., Fogassi, L., Gallese, V., and Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297, 846–848. doi: 10.1126/science.1070311

Kröger, B. J., Birkholz, P., Hoffmann, R., and Meng, H. (2010). "Audiovisual tools for phonetic and articulatory visualization in computer-aided pronunciation training," in *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Vol. 5967*, eds A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt (Dublin: Springer), 337–345.

Kröger, B. J., and Kannampuzha, J. (2008). "A neurofunctional model of speech production including aspects of auditory and audio-visual speech perception," in *International Conference onf Auditory-Visual Speech Processing* (Queensland), 83–88.

Kröger, B. J., Kannampuzha, J., and Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Commun.* 51, 793–809. doi: 10.1016/j.specom.2008.08.002

Kroos, C. (2012). Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). *J. Phonet.* 40, 453–465. doi: 10.1016/j.wocn.2012.03.002

Ladefoged, P., and Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.

Levitt, J. S., and Katz, W. F. (2008). "Augmented visual feedback in second language learning: training Japanese post-alveolar flaps to American English speakers," in *Proceedings of Meetings on Acoustics*, *Vol. 2* (New Orleans, LA), 060002.

Levitt, J. S., and Katz, W. F. (2010). "The effects of EMA-based augmented visual feedback on the English speakers' acquisition of the Japanese flap: a perceptual study," in *Procceedings of Interspeech* (Chiba: Makuhari), 1862–1865.

Liu, X., Hairston, J., Schrier, M., and Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* 35, 1219–1236. doi: 10.1016/j.neubiorev.2010.12.012

Liu, Y., Massaro, D. W., Chen, T. H., Chan, D., and Perfetti, C. (2007). "Using visual speech for training Chinese pronunciation: an in-vivo experiment," in *SLaTE*, 29–32.

Maas, E., Mailend, M.-L., and Guenther, F. H. (2015). Feedforward and feedback control in apraxia of speech (AOS): effects of noise masking on vowel production. *J. Speech Lang. Hear. Res.* 58, 185–200. doi: 10.1044/2014_JSLHR-S-13-0300

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., et al. (2008). Principles of motor learning in treatment of motor speech disorders. *Am. J. Speech Lang. Pathol.* 17, 277–298. doi: 10.1044/1058-0360(2008/025)

Marian, V. (2009). "Audio-visual integration during bilingual language processing," in *The Bilingual Mental Lexicon: Interdisciplinary Approaches,* ed A. Pavlenko (Bristol, UK: Multilingual Matters), 52–78.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Dev.* 55, 1777–1788. doi: 10.2307/1129925

Massaro, D. W. (2003). "A computer-animated tutor for spoken and written language learning," in *Proceedings of the 5th International Conference on Multimodal Interfaces* (New York, NY: ACM), 172–175. doi: 10.1145/958432.958466

Massaro, D. W., Bigler, S., Chen, T. H., Perlman, M., and Ouni, S. (2008). "Pronunciation training: the role of eye and ear," in *Proceedings of Interspeech* (Brisbane, QLD), 2623–2626.

Massaro, D. W., and Cohen, M. M. (1998). "Visible speech and its potential value for speech training for hearing-impaired perceivers," in *STiLL-Speech Technology in Language Learning* (Marholmen), 171–174.

Massaro, D. W., and Light, J. (2003). "Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/," in *Proceedings of Eurospeech (Interspeech)* (Geneva: 8th European Conference on Speech Communication and Technology).

Massaro, D. W., Liu, Y., Chen, T. H., and Perfetti, C. (2006). "A multilingual embodied conversational agent for tutoring speech and language learning," in *Proceedings of Interspeech* (Pittsburgh, PA).

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., and Wallace, M. E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: a theoretical model of stuttering. *Contemp. Issues Commun. Sci. Disord.* 31, 105–122.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Mehta, S., and Katz, W. F. (2015). Articulatory and acoustic correlates of English front vowel productions by native Japanese speakers. *J. Acoust. Soc. Am.* 137, 2380–2380. doi: 10.1121/1.4920648

Mochida, T., Kimura, T., Hiroya, S., Kitagawa, N., Gomi, H., and Kondo, T. (2013). Speech misperception: speaking and seeing interfere differently with hearing. *PLoS ONE* 8:e68619. doi: 10.1371/journal.pone.0068619

Möttönen, R., Schürmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neurosci. Lett.* 363, 112–115. doi: 10.1016/j.neulet.2004.03.076

Navarra, J., and Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Nordberg, A., Göran, C., and Lohmander, A. (2011). Electropalatography in the description and treatment of speech disorders in five children with cerebral palsy. *Clin. Linguist. Phon.* 25, 831–852. doi: 10.3109/02699206.2011.573122

Numbers, M. E., and Hudgins, C. V. (1948). Speech perception in present day education for deaf children. *Volta Rev.* 50, 449–456.

O'Neill, J. J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *J. Speech Hear. Disord.* 19, 429–439.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage* 25, 333–338. doi: 10.1016/j.neuroimage.2004.12.001

Ouni, S. (2013). Tongue control and its implication in pronunciation training. *Comp. Assist. Lang. Learn.* 27, 439–453. doi: 10.1080/09588221.2012.761637

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., and Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Hum. Brain Mapp.* 27, 471–477. doi: 10.1002/hbm.20190

Pochon, J. B., Levy, R., Fossati, P., Lehericy, S., Poline, J. B., Pillon, B., et al. (2002). The neural system that bridges reward and cognition in humans: an fMRI study. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5669–5674. doi: 10.1073/pnas.082111099

Preston, J. L., and Leaman, M. (2014). Ultrasound visual feedback for acquired apraxia of speech: a case report. *Aphasiology* 28, 278–295. doi: 10.1080/02687038.2013.852901

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, 2102–2115. doi: 10.1044/2014_JSLHR-S-14-0031

Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360. doi: 10.1038/nrn2811

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Reetz, H., and Jongman, A. (2009). *Phonetics: Transcription, Production, Acoustics, and Perception.* Chichester: Wiley-Blackwell.

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-reading,* eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum Associates), 97–114.

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194. doi: 10.1016/S0166-2236(98)01260-0

Rizzolatti, G., Cattaneo, L., Fabbri-Destro, M., and Rozzi, S. (2014). Cortical mechanisms underlying the organization of goal-directed actions and mirror neuron-based action understanding. *Physiol. Rev.* 94, 655–706. doi: 10.1152/physrev.00009.2013

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci*. 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Sams, M., Möttönen, R., and Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cogn. Brain Res.* 23, 429–435. doi: 10.1016/j.cogbrainres.2004.11.006

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., and Gracco, V. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Exp. Brain Res.* 227, 275–288. doi: 10.1007/s00221-013-3510-8

Schmidt, R., and Lee, T. (2013). *Motor Learning and Performance: From Principles to Application, 5th Edn*. Champaign, IL: Human Kinetics.

Scruggs, T. E., Mastropieri, M. A., and Casto, G. (1987). The quantitative synthesis of single-subject research methodology and validation. *Remedial Special Educ.* 8, 24–33. doi: 10.1177/074193258700800206

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., and Escobar, C. (1986). Early intervention for children with conduct disorders: a quantitative synthesis of single-subject research. *Behav. Disord.* 11, 260–271.

Shirahige, C., Oki, K., Morimoto, Y., Oisaka, N., and Minagi, S. (2012). Dynamics of posterior tongue during pronunciation and voluntary tongue lift movement in young adults. *J. Oral. Rehabil.* 39, 370–376. doi: 10.1111/j.1365-2842.2011.02283.x

Sigrist, R., Rauter, G., Riener, R., and Wolf, P. (2013). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. *Psychon. Bullet. Rev.* 20, 21–53. doi: 10.3758/s13423-012-0333-8

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2007a). Speech-associated gestures, Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277.

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2006). "Lending a helping hand to hearing: another motor theory of speech perception." in *Action to Language Via the Mirror Neuron System,* ed M. A. Arbib (Cambridge: Cambridge University Press), 250–285.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007b). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Stella, M., Stella, A., Sigona, F., Bernardini, P., Grimaldi, M., and Gili Fivela, B. (2013). "Electromagnetic Articulography with AG500 and AG501," in *14th Annual Conference of the International Speech Communication Association* (Lyon), 1316–1320.

Stevens, K. N. (2008). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N., and Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: a study of retroflex stop consonants. *J. Phonet.* 3, 215–233.

Stevens, K. N., and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoustic. Soc. Am.* 64, 1358–1368. doi: 10.1121/1.382102

Suemitsu, A., Ito, T., and Tiede, M. (2013). An EMA-based articulatory feedback approach to facilitate L2 speech production learning. *J. Acoustic. Soc. Am.* 133, 3336. doi: 10.1121/1.4805613

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoustic. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q., and McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Q. J. Exp. Psychol. Hum. Exp. Psychol.* 36, 51–74. doi: 10.1080/14640748408401503

Swinnen, S. P., Walter, C. B., Lee, T. D., and Serrien, D. J. (1993). Acquiring bimanual skills: contrasting forms of information feedback for interlimb decoupling. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 1328. doi: 10.1037/0278-7393.19.6.1328

Terband, H., and Maassen, B. (2010). Speech motor development in Childhood Apraxia of Speech: generating testable hypotheses by neurocomputational modeling. *Folia Phoniatr. Logop.* 62, 134–142. doi: 10.1159/000287212

Terband, H., Maassen, B., Guenther, F. H., and Brumberg, J. (2009). Computational neural modeling of speech motor control in Childhood Apraxia of Speech (CAS). *J. Speech Lang. Hear. Res.* 52, 1595–1609. doi: 10.1044/1092-4388(2009/07-0283)

Terband, H., Maassen, B., Guenther, F. H., and Brumberg, J. (2014a). Auditory–motor interactions in pediatric motor speech disorders: neurocomputational modeling of disordered development. *J. Communic. Disord.* 47, 17–33. doi: 10.1016/j.jcomdis.2014.01.001

Terband, H., van Brenk, F., and van Doornik-van der Zee, A. (2014b). Auditory feedback perturbation in children with developmental speech sound disorders. *J. Communic. Disord.* 51, 64–77. doi: 10.1016/j.jcomdis.2014.06.009

Tian, X., and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* 1:166. doi: 10.3389/fpsyg.2010.00166

Uddin, L. Q., Molnar-Szakacs, I., Zaidel, E., and Iacoboni, M. (2006). rTMS to the right inferior parietal lobule disrupts self–other discrimination. *Soc. Cogn. Affect. Neurosci.* 1, 65–71. doi: 10.1093/scan/nsl003

Wik, P., and Engwall, O. (2008). "Can visualization of internal articulators support speech perception?," in *Proceedings of Interspeech* (Brisbane), 2627–2630.

Wilson, S. M., and Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage* 33, 316–325. doi: 10.1016/j.neuroimage.2006.05.032

Wilson, S., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Yano, J., Shirahige, C., Oki, K., Oisaka, N., Kumakura, I., Tsubahara, A., et al. (2015). Effect of visual biofeedback of posterior tongue movement on articulation rehabilitation in dysarthria patients. *J. Oral Rehabil.* 42, 571–579. doi: 10.1111/joor.12293

Zaehle, T., Geiser, E., Alter, K., Jancke, L., and Meyer, M. (2008). Segmental processing in the human auditory dorsal stream. *Brain Res.* 1220, 179–190. doi: 10.1016/j.brainres.2007.11.013

Conflict of Interest Statement: This research was partially supported by a grant to Vulintus, LLC entitled "Development of a software package for speech therapy" (NIH-SBIR 1 R43 DC013467). However, the sources of support for this work had no role in the study design, collection, analysis or interpretation of data, or the decision to submit this report for publication. The corresponding author (William F. Katz) had full access to all of the data in the study and takes complete responsibility for the integrity and accuracy of the data. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.