Research Article

Test–Retest Reliability of Behavioral Assays of Feedforward and Feedback Auditory–Motor Control of Voice and Articulation

Mara R. Kapsner-Smith,a Defne Abur,b,c,d Tanya L. Eadie,a and Cara E. Steppb,e,f

a Department of Speech and Hearing Sciences, University of Washington, Seattle b Department of Speech, Language, and Hearing Sciences, Boston University, MA c Department of Computational Linguistics, Center for Language and Cognition, University of Groningen, the Netherlands d Research School of Behavioral and Cognitive Neurosciences, University of Groningen, the Netherlands e Department of Biomedical Engineering, Boston University, MA f Department of Otolaryngology–Head and Neck Surgery, Boston University School of Medicine, MA

ARTICLE INFO

Article History:
Received January 16, 2023
Revision received July 24, 2023
Accepted September 25, 2023
Editor: Raymond D. Kent

https://doi.org/10.1044/2023_JSLHR-23-00038

ABSTRACT

Purpose: Behavioral assays of feedforward and feedback auditory–motor control of voice and articulation frequently are used to make inferences about underlying neural mechanisms and to study speech development and disorders. However, no studies have examined the test–retest reliability of such measures, which is critical for rigorous study of auditory–motor control. Thus, the purpose of the present study was to assess the reliability of assays of feedforward and feedback control in voice versus articulation domains.

Method: Twenty-eight participants (14 cisgender women, 12 cisgender men, one transgender man, one transmasculine/nonbinary) who denied any history of speech, hearing, or neurological impairment were measured for responses to predictable versus unexpected auditory feedback perturbations of vocal (fundamental frequency, \(f_o\)) and articulatory (first formant, \(F_1\)) acoustic parameters twice, with 3–6 weeks between sessions. Reliability was measured with intraclass correlations.

Results: Opposite patterns of reliability were observed for \(f_o\) and \(F_1\); \(f_o\) reflexive responses showed good reliability and \(f_o\) adaptive responses showed poor reliability, whereas \(F_1\) reflexive responses showed poor reliability and \(F_1\) adaptive responses showed moderate reliability. However, a criterion-referenced categorical measurement of \(f_o\) adaptive responses as typical versus atypical showed substantial test–retest agreement.

Conclusions: Individual responses to some behavioral assays of auditory–motor control of speech should be interpreted with caution, which has implications for several fields of research. Additional research is needed to establish reliable criterion-referenced measures of \(F_1\) adaptive responses as well as \(f_o\) and \(F_1\) reflexive responses. Furthermore, the opposite patterns of test–retest reliability observed for voice versus articulation add to growing evidence for differences in underlying neural control mechanisms.

Humans learn and maintain movement patterns for intelligible speech by monitoring sensory feedback, especially in the auditory domain. By comparing auditory feedback of speech with internal predictions, speakers can detect errors and produce online corrections as well as updates to stored motor programs (Tourville & Guenther, 2011). In studies of sensorimotor control of voice and articulation, experimenters often measure behavioral responses to altered auditory feedback to assess auditory–motor control of speech (e.g., Burnett et al., 1997; Cai et al., 2011; Daliri et al., 2020; Houde & Jordan, 1998; Jones & Munhall, 2000; Larson & Robin, 2016; Lester-Smith et al., 2020; Villacorta et al., 2007). These behavioral assays provide experimental data that have been used to develop and test neurocomputational models of speech.
sensorimotor control and are used in combination with neurophysiological methods to investigate the neurobiology of speech (e.g., Houde & Chang, 2015; Tourville & Guenther, 2011). Studies of children and aging adults have contributed to our understanding of speech development across the life span (Caudrelier & Rochet-Capellan, 2019; H. Liu, Russo, & Larson, 2010; Scheerer et al., 2020). Recently, an increasing number of studies use these paradigms to examine impaired sensorimotor control in clinical populations (Caudrelier & Rochet-Capellan, 2019; Weerathunge, Tomassi, & Stepp, 2022). Such experiments can elucidate the pathophysiology of speech disorders as well as test hypotheses about sensorimotor control of speech using clinical populations as a model. However, to date, no studies have directly examined the reliability of behavioral assays of auditory–motor control of voice and articulation. This information is crucial to appropriately interpret differences between and within experimental groups in participants’ responses to altered auditory feedback.

**Behavioral Assays of Feedforward and Feedback Auditory–Motor Control**

Sensorimotor control of speech is an incredibly complex task, requiring precise control of approximately 100 muscles to produce rapid movements and quickly changing acoustic features, on a scale of 50–300 ms (Parrell et al., 2019). At the same time, a speaker must be able to adapt to changing environmental conditions, such as background noise, to maintain intelligibility. This is achieved through a combination of feedforward and feedback control mechanisms. Feedforward control involves execution of preprogrammed motor commands for speech movements that allow for fast and fluent production. At the same time, an internal prediction of the sensory outcomes of the movements is activated. Comparison of this prediction to the actual sensory feedback allows the speaker to detect errors and generate corrective, feedback-based motor commands. Persistent error signals may lead to changes to feedforward motor commands, that is, sensorimotor learning.

An extensive body of literature has explored the role of auditory feedback in sensorimotor control of speech by applying near–real-time perturbations to parameters of auditory feedback. These include parameters related to voice, such as fundamental frequency ($f_0$; e.g., Burnett et al., 1997), and parameters related to articulation, such as formants (e.g., Houde & Jordan, 1998). When a speaker is exposed to persistent, predictable perturbations of their auditory feedback over many trials, they typically will adapt by progressively opposing the change across trials and will demonstrate a brief period of persistence of this change after the perturbation is removed (Daliri & Dittman, 2019; Houde & Jordan, 1998; Jones & Munhall, 2000; Purcell & Munhall, 2006a; Villacorta et al., 2007). This is interpreted as reflecting updates to underlying feedforward motor commands based on an error signal, often referred to as an adaptive response (Parrell & Houde, 2019; Tourville & Guenther, 2011). When a speaker is exposed to an unexpected, sudden-onset perturbation of their auditory feedback, they typically will respond with a rapid correction opposing the direction of the perturbation (Cai et al., 2011; Daliri et al., 2020; Larson & Robin, 2016; Niziolet & Guenther, 2013; Purcell & Munhall, 2006b). This is interpreted as a reflex-like correction produced by feedback control loops (Parrell & Houde, 2019; Tourville & Guenther, 2011). Measures such as the magnitude of the response to predictable or unexpected auditory feedback perturbations are thus used as behavioral assays of underlying feedforward and feedback control mechanisms, respectively, although both mechanisms are involved in these tasks.

**Use of Assays to Study Neurophysiology of Speech**

Behavioral assays of speech auditory–motor control have been combined with neurophysiological methods to elucidate underlying control mechanisms. Feedforward control assays combined with noninvasive neurostimulation methods including anodal transcranial direct stimulation and repetitive transcranial magnetic stimulation have provided information about the neurophysiological bases of speech sensorimotor learning. To date, these studies have focused on adaptive responses to predictable perturbations of the first formant ($F_1$), an acoustic parameter related to articulatory movements of the tongue and jaw. Changes in the magnitude and/or rate of adaptation in response to stimulation support the role of a sensorimotor corticocerebellar loop in integrating auditory feedback error signals into feedforward motor plans, that is, adaptive auditory–motor learning for speech (Deroche et al., 2017; Lametti et al., 2018; Scott et al., 2020; Shum et al., 2011; Tang et al., 2021).

Feedback control assays combined with methods including positron emission tomography, magnetoencephalography, electroencephalography, electrocorticography, and functional magnetic resonance imaging (fMRI) have characterized the role of the auditory cortex in comparison of predicted and actual auditory feedback for voice and articulation. Responses of the auditory cortex are significantly smaller when speakers hear their own speech during speech production (i.e., auditory feedback) than when they listen passively to playback of the same speech signal or when they hear auditory feedback that has been
altered, a phenomenon known as speaking-induced suppression (Chang et al., 2013; Flinker et al., 2010; Greenlee et al., 2011; Hirano et al., 1997; Houde et al., 2002; Sato & Shiller, 2018). This evidence supports the hypothesis that auditory feedback is compared to an internal prediction to detect speech errors. Further evidence comes from studies of responses to altered auditory feedback, which reveal enhanced auditory cortical responses to altered auditory feedback during speech compared to passive listening to playback of the same altered signals, known as speech perturbation response enhancement (Behroozmand et al., 2009; Chang et al., 2013; Greenlee et al., 2013; Kort et al., 2014; Parkinson et al., 2012). Furthermore, studies using fMRI during feedback control assays show increased activity in and connectivity between bilateral posterior superior temporal cortices, supporting their role in error detection, and right and/or bilateral motor cortices, suggesting a role in translating error signals into corrective motor output (Behroozmand et al., 2015; Floegel et al., 2020; Niziolek & Guenther, 2013; Tourville et al., 2008). Although such studies have contributed substantially to understanding the neurobiology of speech sensorimotor control, information about the test–retest reliability of the behavioral measures is crucial to understand whether these responses reflect state (temporary, situational) versus trait (stable, long-lasting) characteristics of the speaker. Furthermore, information about reliability could provide future directions for mechanistic studies to determine what causes the behavior of an individual to change.

Use of Assays to Study Speech Development

Studies of the developmental trajectory of auditory feedback perturbation responses suggest that young children use auditory feedback to guide sensorimotor control and learning for both vocal and articulatory parameters. Children as young as 24 months have shown increased responses to unpredictable vocal parameter perturbations (H. Liu, Russo, & Larson, 2010; P. Liu, Chen, et al., 2010; MacDonald et al., 2012; Scheerer et al., 2013, 2016, 2020; van BRENK & Terband, 2020). This may be due to neurodevelopmental factors such as myelination that increase neural transmission speed and reduce processing and response times for auditory feedback responses.

Two studies have investigated the influence of aging on responses to auditory feedback perturbations of $f_0$. Aging appears to increase the magnitude of compensatory responses to unpredictable $f_0$ perturbations, but it is unknown whether this reflects changes in feedback control mechanisms or physiological pitch control more generally (H. Liu, Russo, & Larson, 2010; P. Liu et al., 2011). The influence of aging on responses to auditory feedback perturbations of articulatory parameters and on adaptive responses to predictable perturbations remains unexplored.

Use of Assays to Study Speech Disorders

An increasing number of auditory feedback perturbation studies examine clinical populations. One purpose of such studies is to elucidate the pathophysiology of voice and articulation disorders that previously were poorly understood, such as developmental stuttering and hyperfunctional voice disorders. Numerous studies document a broad, likely developmental impairment of sensorimotor control and learning in individuals who stutter, including articulatory (Cai et al., 2012, 2014; Daliri et al., 2018; Kim, Daliri, et al., 2020; Sengupta et al., 2016), vocal (Bauer et al., 2007; Loucks et al., 2012; Sares et al., 2018, 2020), and limb control (Kim, Daliri, et al., 2020). Recent evidence points to impaired adaptive auditory–motor learning for voice in some people with hyperfunctional voice disorders, with likely preserved feedback auditory–motor control (Abur, Subaciute, Kapsner-Smith, et al., 2021; Stepp et al., 2017). Such studies may have substantial implications for diagnosis and treatment of speech disorders.

In contrast, patient populations with known neurogenic impairments can serve as models to understand underlying neural mechanisms of typical sensorimotor control. For example, in people with aphasia secondary to left-hemisphere stroke, lesion mapping, combined with feedback control assays and neurophysiological methods, has supported the role of frontotemporal networks in error detection and correction (Behroozmand et al., 2018, 2022). Behavioral and electroencephalographical data from people with cerebellar disease suggest cerebellar function is important for feedforward auditory–motor adaptive learning and may play a role in feedback auditory–motor control through interactions with cortical regions that modulate auditory–motor integration (Hilger, 2020; Houde et al., 2019; Li et al., 2019; Parrell et al., 2017). Finally, people with Parkinson’s disease (PD), a neurodegenerative disease associated with impaired dopamine production, have shown increased responses to feedback control assays for the vocal parameters $f_0$ and intensity while off dopaminergic medications (X. Chen et al., 2013; H. Liu et al., 2012; Mollaei et al., 2016) but
decreased responses for the articulatory parameter \( F_1 \) (Mollaei et al., 2016). This dissociation of impaired feedback auditory–motor control of voice versus articulation suggests differences in underlying neural mechanisms. Although no studies to date have examined differences in medication state within the same individuals, studies of people with PD while on dopaminergic medications have shown typical feedback control responses for both vocal and articulatory parameters (Abur, Subaciute, Daliri, et al., 2021; Kiran & Larson, 2001), suggesting that dopaminergic signaling is involved in typical auditory–motor feedback control. Crucially, most of these studies of clinical populations have relied on group-level findings, despite substantial variability in behavior at the individual level (e.g., Abur, Subaciute, Daliri, et al., 2021). Information about the test–retest reliability of individual responses to behavioral assays of speech sensorimotor control will help determine whether they are rigorous measures capable of reliably detecting differences between groups or changes in individuals over time.

**Reliability of Behavioral Assays of Auditory–Motor Control**

Reliability refers to “the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials” (Merriam-Webster, n.d.). Good test–retest reliability suggests that a measure is an accurate representation of an individual’s performance rather than artifact related to irrelevant aspects of the testing session or environment. Stability of a measure over time may also suggest that it is sensitive to trait, rather than state, characteristics of the individual being measured, for example, a stable capacity of the speech sensorimotor control system versus one that fluctuates in response to internal and/or external conditions. Despite the importance of reliability for interpretation of results, to date, no studies have explicitly examined the test–retest reliability of behavioral assays of auditory–motor control of speech. One study examined adaptive auditory–motor learning for voice in response to two differently sized perturbations of \( f_s \) — one standard perturbation and the other that was personalized to the individual’s \( f_s \) discrimination threshold (Alemi et al., 2020). Participants’ responses were only weakly related across the two experiments, with substantial variability in the magnitude and the direction of \( f_s \) changes. Another study examined feedback control responses to perturbations of \( F_1 \) with and without the addition of noise (i.e., artificially increased variability) in participants’ auditory feedback prior to the experiment (Niziolek & Parrell, 2021). The authors noted that interpretation of results in this study was limited by a high degree of intra-individual variability, potentially due to “limited stability of auditory feedback compensation measures... across sessions” (Niziolek & Parrell, 2021, p. 2169). In addition, it is unknown whether learning effects occur with repeated exposure to identical tasks. Although learning effects have been observed in studies of reaching movements, these studies often involve longer exposure to perturbations (e.g., Shadmehr & Holcomb, 1997), and important differences have been documented between learning in limb and speech systems, such as the presence (limb) and absence (speech) of explicit learning mechanisms (Kim & Max, 2020; Lametti et al., 2020). These studies highlight the importance of directly investigating the test–retest reliability of measures of feedforward and feedback auditory–motor control of voice and articulation.

To interpret within-group and between-groups differences in auditory perturbation experiments, investigators must know how reliable responses to auditory perturbation are in individuals with typical speech. To date, no studies have examined the test–retest reliability of common assays of feedforward and feedback control of voice and articulation. Thus, the purpose of the present study was to measure the test–retest reliability of adaptive responses to predictable perturbations and reflexive responses to unexpected perturbations in both the voice \( (f_v) \) and articulation \( (F_1) \) domains. Given the variability of responses to auditory feedback perturbations between individuals and across similar experiments (Alemi et al., 2020; Niziolek & Parrell, 2021), we hypothesized that test–retest reliability of these assays may be only moderate. As a secondary analysis, we also examined whether systematic learning effects occurred with repeated exposure to typical experimental tasks measured several weeks apart.

**Method**

All procedures were approved by the institutional review board of the University of Washington.

**Participants**

Twenty-eight adults reporting normal speech, language, and hearing function were recruited to participate in the study (14 cisgender women, 12 cisgender men, one transgender man, one transmasculine/nonbinary; 16 assigned female at birth, 12 assigned male at birth; \( M_{age} = 25 \) years, \( SD = 4.5 \), range: 18–35 years). The sample size provided 80% power to detect medium effect sizes (intraclass correlation coefficient [ICC] ≥ .6). Individuals with a self-reported history of neurological disease, stroke, or any health conditions affecting communication or the ability to pay attention were excluded. Individuals who were smokers were excluded due to potential effects on vocal
function. Individuals who reported formal singing training or were active vocal performers were excluded. Speakers of tonal languages were excluded. All participants passed a hearing screening with thresholds at or below 25 dB HL at 125, 250, 500, 1000, 4000, and 8000 Hz (American Speech-Language-Hearing Association, 2018). All participants learned English in infancy.

Participants were screened for voice concerns via interview and administration of the Voice-Related Quality of Life (Hogikyan & Sethuraman, 1999). Participants reported no voice overuse or upper respiratory symptoms for at least 48 hr prior to each recording session. Participants reporting illness or voice overuse were rescheduled to allow for vocal recovery (Hunter & Titze, 2009).

**Procedure**

Adaptive and reflexive responses to $f_0$ and $F_1$ perturbations (Lester-Smith et al., 2020) were each measured twice for all participants. Baseline measures of $f_0$ and $F_1$ control occurred in separate sessions due to additional data collection for a different study and were conducted within a span of 2 weeks (hereafter referred to as Time 1). Retest measures of $f_0$ and $F_1$ control occurred in a single session, which was conducted 3–6 weeks after the second baseline session (hereafter referred to as Time 2). Each session lasted 1–2 hr. We chose a timeline that we expected was unlikely to induce learning and might occur in a longitudinal study design, while also avoiding changes in vocal function over time due to seasonal or other factors. The order of the first two sessions and the order of $f_0$ and $F_1$ tasks in the third session were counterbalanced across participants to control for order effects.

Participants were seated comfortably in front of a monitor that presented visual stimuli. They were provided with water before and during the experiments. Participants wore a head-mounted Shure Omni Mic MX153 7 cm from the center of the mouth at an approximately 45° angle and Etymotic ER-2 insert earphones. The microphone signal was amplified +5 dB and played back to participants via the headphones with a total processing delay of < 35 ms (Kim, Wang, & Max, 2020; Weerathunge et al., 2020). Both the microphone and headphone digitized signals were recorded at 16 kHz for $F_1$ experiments and 44.1 kHz for $f_0$ experiments.

For $f_0$ perturbation experiments, the microphone signal was amplified by an RME QuadMic II Microphone Preamp and digitized by an RME Fireface sound card. Manipulations of $f_0$ were performed by an Eventide Eclipse V4. The manipulated signal was amplified by a Behringer Xenyx Q802USB amplifier and played back to the participant via headphones, as well as being digitized via the RME Fireface sound card. Initiation of trials and Eventide Eclipse manipulations was controlled by a custom MATLAB program (MathWorks, 2016).

For $F_1$ perturbation experiments, the microphone signal was amplified and digitized by a MOTU MicroBook IIC microphone preamplifier using MicroBook II CueMix software. $F_1$ manipulations were performed by Audapter (Cai et al., 2008), a MATLAB software package that performs real-time transformations of acoustic parameters. The signal was then sent to headphones via a Behringer Xenyx Q802USB amplifier. Initiation of trials and Audapter manipulations were controlled by a custom MATLAB program (MathWorks, 2016).

There were three $f_0$ perturbation conditions: adaptation shift-up, adaptation control, and reflex shift-up (10 min each; Lester-Smith et al., 2020). Adaptation conditions always occurred before the reflex condition. The order of the two adaptation conditions was counterbalanced across participants. Each condition consisted of 108 trials, during which participants produced a 2-s sustained /a/. During the adaptation shift-up condition, participants received unperturbed feedback during the first 24 trials (baseline phase), gradually upward-shifted feedback of +3.4 cents per trial relative to the participant’s $f_0$ during the next 30 trials (ramp phase), 30 trials at the maximum perturbation of +100 cents (hold phase), and 24 trials of unperturbed feedback (aftereffect phase). Perturbations during the adaptation condition, when applicable, were applied for the duration of the trial. During the adaptation control condition, participants received unperturbed feedback for all 108 trials. During the reflex shift-up condition, participants produced a 2-s sustained /a/ for 12 blocks of nine trials (total of 108). Each block included two perturbed trials. Perturbations consisted of a sudden +100 cent shift of $f_0$. Onset of perturbations was jittered between 0.5 and 1 s after the onset of phonation. After onset, perturbations remained on for the remainder of the trial.

There were two $F_1$ perturbation conditions: adaptation shift-up and reflex shift-up (10 min each; Lester-Smith et al., 2020). The adaptation condition always occurred before the reflex condition. Each condition consisted of 108 trials. Participants produced a prolonged word for approximately 2 s when presented with the word on the screen. During the adaptation shift-up condition, a block consisted of one presentation of each of the stimuli “bid,” “tid,” and “hid,” in random order (36 blocks × 3 trials = 108 trials). These stimuli were chosen so that they would be perceived as a real word when $F_1$ was perturbed upward (toward “bed,” “Ted,” and “head”). Participants received unperturbed feedback during the first eight blocks (baseline phase), gradually upward-shifted $F_1$ feedback of
+1.03% per trial relative to \( F_1 \) produced by the participant across the next 10 blocks (ramp phase), 10 blocks at the maximum perturbation of +30% of \( F_1 \) (hold phase), and eight blocks of unperturbed feedback (aftereffect phase).

During the reflex shift-up condition, a block consisted of three presentations of each of the stimuli “bid,” “tid,” and “hid,” in random order (12 blocks × 9 trials = 108 trials). Each block included two perturbed trials. Perturbations consisted of a +30% shift-up of \( F_1 \). For both the adaptation and reflex conditions, perturbations were applied for the duration of the trial, beginning with the onset of speech. Schematics for adaptation and reflex experiments are provided in Figure 1.

**Data Analysis**

For adaptation measures, the mean value of \( f_o \) or \( F_1 \) was calculated during the window 40–120 ms after the onset of the vowel in each trial. This window was chosen to assess contributions of the feedforward sensorimotor control system after the voice stabilized but before auditory feedback responses are expected to occur (Burnett et al., 1998). For the \( f_o \) adaptation experiment, mean \( f_o \) of each trial was extracted using an autocorrelation method via Praat (Boersma & Weenink, 2016) and converted to cents, a relative measure of \( f_o \), by normalizing to the average of the 24 baseline trials. Trials with inadequate pitch tracking or duration were replaced with the average of the trial before and the trial after. For \( f_o \), a total of 0.36% of all trials were replaced, averaging 0.4 trials per experimental condition (range: 0–9 trials per condition). The \( f_o \) shift-up condition was then normalized to the \( f_o \) control condition by subtracting the mean \( f_o \) in cents of the respective control trial from each shift-up trial. This normalization accounted for any change in \( f_o \) across trials that was not related to the auditory feedback perturbation. This procedure was used for \( f_o \) and not \( F_1 \) because \( f_o \) tends to drift upward over many trials (Jones & Munhall, 2000). For the \( F_1 \) adaptation experiment, \( F_1 \) was extracted using linear predictive coding and converted to a percentage relative to the mean \( F_1 \) of the 24 baseline trials (Lester-Smith et al., 2020). For \( F_1 \), no trials needed to be replaced. For both experiments, the average of the 30 trials in the hold phase was used for statistical analyses of the adaptive responses. A second measure using the mean of the last 15 trials (second half) of the hold phase was also calculated and compared to the primary measure. This was conducted as a post hoc analysis to assess whether responses were more stable in the later portion of the hold phase, for example, if adaptation has plateaued.

For reflex measures, \( f_o \) or \( F_1 \) traces were extracted for all trials. Perturbed \( f_o \) trials were normalized in cents relative to the 100-ms period preceding the start of the perturbation. Perturbed \( F_1 \) trials were normalized as a percentage relative to the mean trajectory of unperturbed trials containing the same word (Lester-Smith et al., 2020). Perturbed trials with inadequate pitch tracking or duration

---

**Figure 1.** Schematics of (a) adaptation conditions and of a single perturbed reflex trial for (b) \( f_o \) and (c) \( F_1 \). Dashed lines represent the perturbation magnitude, and shaded regions represent the analysis windows. \( f_o \) = fundamental frequency; \( F_1 \) = first formant.
were removed from the analysis. For $f_o$, there was an average of 22.89 usable reflex trials per participant (range: 17–24 out of 24 possible trials). For $F_1$, only two reflex trials were removed from the entire data set, resulting in an average of 23.99 usable $F_1$ reflex trials per participant (range: 23–24 out of 24 possible trials). The mean normalized value of $f_o$ or $F_1$ was calculated across all perturbed trials during the window 120–240 ms after the onset of the perturbation (Lester-Smith et al., 2020), consistent with timing of reflexive responses reported by Tourville et al. (2008) and Hain et al. (2000). This mean was used for statistical analyses of the reflexive responses.

### Statistical Analyses

For $f_o$, adaptation, $F_1$ adaptation, $f_o$ reflex, and $F_1$ reflex, ICCs were used to assess the reliability of the responses at Time 1 and Time 2. ICCs and 95% confidence intervals were calculated using SPSS (Version 26, IBM, Inc.) based on single measures, absolute agreement, and two-way random effects (McGraw & Wong, 1996). Categories of poor (< .50), moderate (.50–.75), good (.76–.90), and excellent (> .90) reliability have been suggested for ICCs (Koo & Li, 2016). As a secondary analysis, two-tailed paired $t$ tests were used to assess whether learning effects occurred from repeated exposure to the behavioral assays, with an $\alpha$ level of .05.

Prior studies of $f_o$ adaptive responses in individuals with typical voice and speech have documented substantial variability, including both responses that oppose the perturbation and responses that follow (i.e., change in the same direction as) the perturbation (e.g., Abur, Subaciute, Kapsner-Smith, et al., 2021; Lester-Smith et al., 2020). An additional categorical analysis of $f_o$ adaptive responses using cutoff criteria from a prior study was therefore conducted to assess its reliability. This analysis was not conducted for the other behavioral assays ($f_o$ reflex, $F_1$ adaptation, $F_1$ reflex) because similar criterion cutoff scores were not available. The mean $f_o$ adaptive response was categorized as typical or atypical based on 90th percentile z-score cutoffs taken from nonsinging control participants in the study of Abur, Subaciute, Kapsner-Smith, et al. (2021). Cohen’s kappa statistic was calculated to assess agreement of this categorical score between Time 1 and Time 2.

### Results

Descriptive statistics for group results for each of the four experiments ($f_o$ adaptation, $F_1$ adaptation, $f_o$ reflex, and $F_1$ reflex) are provided in Table 1. On average, participants opposed the $f_o$ and $F_1$ auditory perturbations at both time points. The average magnitudes of $f_o$ adaptive responses were $-37.7$ ($SD = 70.3$) and $-51.3$ ($SD = 72.3$) with a similar number of participants producing following responses at each time point. The average magnitudes of $f_o$ reflexive responses were $-8.9$ ($SD = 8.7$) and $-10.8$ ($SD = 10.0$). The average magnitudes of $F_1$ adaptive responses were $-9.0$ ($SD = 7.3$) and $-6.6$ ($SD = 5.7$). The average magnitudes of $F_1$ reflexive responses were $-1.1$ ($SD = 1.6$) and $-1.2$ ($SD = 1.6$). Individual data are presented in Figures 2–5. As a secondary analysis, paired $t$ tests were calculated for each of the four experiments to determine if systematic learning effects occurred. There was no significant difference for any of the measures at the two time points ($t$ statistics and $p$ values are provided in Table 1).

For the $f_o$ reflex task, at Time 1, participants responded to an average of 54.7% of perturbed trials in a compensatory direction ($SD = 17.8\%$) and 20.0% of trials in a following direction ($SD = 12.0\%$) and did not respond to 25.3% ($SD = 11.8\%$). Similarly, at Time 2, participants responded to an average of 55.4% of perturbed trials in a compensatory direction ($SD = 18.6\%$) and 19.7% in a following direction ($SD = 14.3\%$) and did not respond to 25.0% ($SD = 11.8\%$).

### Table 1. Descriptive statistics for the magnitude of adaptive and reflexive responses and paired $t$ tests.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Time</th>
<th>Mean response</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>$t(27)$</th>
<th>$p$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_o$ adaptation</td>
<td>1</td>
<td>$-37.7$ cents</td>
<td>$-154.1$</td>
<td>109.4</td>
<td>70.3</td>
<td>.070</td>
<td>.492</td>
<td>0.13</td>
</tr>
<tr>
<td>$f_o$ adaptation</td>
<td>2</td>
<td>$-51.3$ cents</td>
<td>$-233.5$</td>
<td>95.2</td>
<td>72.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_1$ adaptation</td>
<td>1</td>
<td>$-9.0%$</td>
<td>$-19.2$</td>
<td>15.4</td>
<td>7.3</td>
<td>$-2.04$</td>
<td>.051</td>
<td>0.39</td>
</tr>
<tr>
<td>$F_1$ adaptation</td>
<td>2</td>
<td>$-6.6%$</td>
<td>$-14.4$</td>
<td>7.7</td>
<td>5.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_o$ reflex</td>
<td>1</td>
<td>$-8.9$ cents</td>
<td>$-36.4$</td>
<td>5.3</td>
<td>8.7</td>
<td>1.85</td>
<td>.075</td>
<td>0.35</td>
</tr>
<tr>
<td>$f_o$ reflex</td>
<td>2</td>
<td>$-10.8$ cents</td>
<td>$-39.4$</td>
<td>9.8</td>
<td>10.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_1$ reflex</td>
<td>1</td>
<td>$-1.1%$</td>
<td>$-5.6$</td>
<td>1.3</td>
<td>1.6</td>
<td>0.09</td>
<td>.932</td>
<td>0.02</td>
</tr>
<tr>
<td>$F_1$ reflex</td>
<td>2</td>
<td>$-1.2%$</td>
<td>$-5.3$</td>
<td>2.1</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Negative values reflect opposing responses, whereas positive values reflect following responses. $f_o$ = fundamental frequency; $F_1$ = first formant.
ICCs were calculated for each of the four experiments. For $f_o$ adaptive responses, ICC(2, 1) = $-0.052 \, [-0.422, 0.326]$, $p = 0.604$, consistent with poor test–retest reliability. For $F_1$ adaptive responses, ICC(2, 1) = $0.536 \, [0.219, 0.753]$, $p = 0.001$, consistent with moderate test–retest reliability. For $f_o$ reflexive responses, ICC(2, 1) = $0.833 \, [0.663, 0.919]$, $p < 0.001$, consistent with good test–retest reliability. For $F_1$ reflexive responses, ICC(2, 1) = $0.474 \, [0.122, 0.718]$, $p = 0.005$, consistent with poor test–retest reliability.

ICCs were also calculated for the two adaptation experiments using the mean of the last 15 trials (second half) of the hold phase. For $f_o$ adaptive responses, ICC(2, 1) = $-0.115 \, [-0.475, 0.270]$, $p = 0.719$, consistent with poor test–retest reliability. For $F_1$ adaptive responses, ICC(2, 1) = $0.525 \, [0.206, 0.745]$, $p = 0.001$, consistent with moderate test–retest reliability.
test–retest reliability. These results were slightly less reliable than ICCs calculated using the mean of the entire hold phase.

Criterion-referenced assessments are a common way to categorize responses into typical or atypical categories based on normative data. This type of approach was used by Abur, Subaciute, Kapsner-Smith, et al. (2021) to identify atypical $f_o$ adaptive responses in a subgroup of people with vocal hyperfunction. To assess the stability of $f_o$ adaptive responses within a typical range in the present study, participants’ responses were categorized into typical versus atypical using 90th percentile $z$ scores from data collected by Abur, Subaciute, Kapsner-Smith, et al. (2021). Cutoff scores were calculated based on nonsingers with typical voices ($n = 33$). $f_o$ adaptive responses were categorized as typical if they were greater than $-199.5$ cents or less than 97.8 cents. Based on these cutoffs, 26 participants in the present study were categorized as typical responders at both Time 1 and Time 2, one participant was categorized as an atypical responder at both Time 1 and Time 2, and one participant was categorized as an atypical responder at Time 1 and a typical responder at Time 2. This results in 96.4% agreement with $k = .65$ reflecting substantial agreement.

**Discussion**

Test–retest reliability was assessed for common behavioral assays of feedforward and feedback auditory–motor control of voice and articulation in people with typical speech. These included adaptive learning responses and reflexive responses to predictable versus unexpected perturbation of voice ($f_o$) and articulation ($F_1$) parameters of auditory feedback. The magnitudes of $F_1$ adaptive and reflexive responses were similar to those reported by Lester-Smith et al. (2020), which used the same methods. Consistent with our hypothesis, test–retest reliability for the magnitude of $F_1$ adaptive responses was moderate, and reliability for $F_1$ reflexive responses was poor. However, test–retest reliability for the magnitude of $f_o$ adaptive and reflexive responses followed an opposite pattern, with good reliability for reflexive responses and poor reliability for adaptive responses. Notably, when $f_o$ adaptive responses were categorized into typical versus atypical scores based on the distribution of scores among control participants in a prior study (Abur, Subaciute, Kapsner-Smith, et al., 2021), there was substantial agreement between scores at the two time points.

Secondary testing revealed no significant changes from Time 1 to Time 2 suggestive of systematic learning effects in any of the auditory–motor control assays. Although results were borderline for two of the assays ($F_1$ adaptation and $f_o$ reflex), effect sizes were small. This suggests that these assays may be used for repeated testing in study designs with exposure duration and spacing similar to the present study, without substantial concern for systematic changes such as learning effects due to the measurement task itself. However, reliability of these measures should be taken into account when interpreting repeated measures. Concerns regarding reliability may be particularly relevant for studies examining individual responses as well as individual changes in longitudinal and interventional designs. Such studies are crucial for characterization of causal mechanisms and for clinical applications in people with communication disorders.

The opposite patterns in reliability we observed for behavioral assays of auditory–motor control of voice versus articulation suggest differences in underlying neurophysiological mechanisms. For voice, reflexive responses to unpredictable $f_o$ perturbations were highly reliable, whereas the magnitude of adaptive responses to predictable $f_o$ perturbations showed no relationship between time points. It has been suggested that online feedback-based control of parameters related to voice, such as $f_o$ and intensity, may be important for maintaining intelligibility in response to changing environmental or other conditions (Perkell et al., 1997). Vocal parameters may be controlled at multiple time scales, with a phrase component extending across numerous phoneme and word boundaries (Ladd, 2008). Fast and slow components of $f_o$ control may be modeled separately (Fujisaki, 2004), and there is evidence for distinct neural control mechanisms (Dichter et al., 2018). The slower components of vocal parameters thus may be amenable to feedback-based control. In contrast, acoustic parameters associated with speech articulation, such as formants, change more rapidly than the time that is required for online feedback-based responses. Thus, to produce fluent and intelligible speech, a speaker must rely on an internal model for feedforward control of articulation. Theories of speech motor control include auditory feedback as crucial for both learning and maintaining such models (Parrell et al., 2019; Perkell et al., 1997; Tourville & Guenther, 2011). These views are consistent with our findings of more reliable reflexive responses to $f_o$ perturbations than $F_1$, and the reverse finding for adaptive responses.

The present study adds to growing evidence for differences in sensorimotor control mechanisms underlying voice versus articulation. Studies that have compared adaptive and reflexive responses to perturbations of $f_o$ and formants suggest that there is no significant relationship between responses to $f_o$ and formant feedback perturbations in the same speakers (MacDonald & Munhall, 2012; Weerathunge, Voon, et al., 2022). Average compensation for unexpected perturbations of $f_o$ (e.g., Burnett et al.,...
The magnitude of adaptive responses to persistent $f_o$ perturbations had poor test–retest reliability in the present study. This may reflect susceptibility of the adaptive auditory–motor control system for voice to factors that were not controlled. People with typical speech display a wide range of $f_o$ adaptive responses, including large compensatory responses as well as following responses, that is, changes in $f_o$ in the same direction as the feedback perturbation (e.g., Abur, Subaciute, Kapsner-Smith, et al., 2021; Lester-Smith et al., 2020). Researchers have suggested that the $f_o$ auditory–motor control mechanism may respond flexibly to auditory feedback errors depending on whether they are perceived as internally versus externally generated (Burnett et al., 1998) and that there may be separate compensatory versus imitative modes resulting from different underlying neural processes (Korzyukov et al., 2012; Li et al., 2013; Patel et al., 2014, 2019). However, to date, these studies have focused on reflexive responses to unpredictable $f_o$ perturbations. When averaged across trials, these responses were highly reliable in the present study. It is possible that $f_o$ adaptive responses may have similar flexibility and may vary depending on uncontrolled factors such as internal states of the participant during an individual testing session.

Although the magnitude of $f_o$ adaptive responses was not reliable across testing sessions, categorical scoring of $f_o$ adaptive responses as typical versus atypical showed substantial test–retest agreement. This scoring method was implemented by Abur, Subaciute, Kapsner-Smith, et al. (2021) to detect differences in people with and without hyperfunctional voice disorders that were masked in comparisons of group means due to the presence of atypical responses at both ends of the distribution (i.e., compensating vs. following). Because there is substantial variability in $f_o$ adaptive responses in people with typical speech, scores that fall significantly outside of that distribution are remarkable. This scoring method has relevance for studying impaired auditory–motor control of speech in people with communication disorders. Studies with larger normative samples and/or meta-analysis of existing $f_o$ adaptation studies are needed to establish more precise cutoff scores for typical $f_o$ adaptive responses. Furthermore, normative data for $F_1$ reflexive responses could be used to assess the reliability of categorical measurement of articulatory feedback control using a similar approach.

One limitation of the present study is the difference in tasks (sustained phonation vs. words) for the $f_o$ versus $F_1$ experiments. This approach was used because most current studies of $f_o$ adaptive responses use sustained phonation, and most current studies of $F_1$ adaptive responses use speech or speechlike tasks. It is possible that participants’ $f_o$ feedforward targets may be less well defined for sustained phonation than their $F_1$ feedforward targets.

The magnitude of adaptive responses to persistent $f_o$ perturbations had poor test–retest reliability in the present study. This may reflect susceptibility of the adaptive auditory–motor learning system for voice to factors that were not controlled. People with typical speech display a wide range of $f_o$ adaptive responses, including large compensatory responses as well as following responses, that is, changes in $f_o$ in the same direction as the feedback perturbation (e.g., Abur, Subaciute, Kapsner-Smith, et al., 2021; Lester-Smith et al., 2020). Researchers have suggested that the $f_o$ auditory–motor control mechanism may respond flexibly to auditory feedback errors depending on whether they are perceived as internally versus externally generated (Burnett et al., 1998) and that there may be separate compensatory versus imitative modes resulting from different underlying neural processes (Korzyukov et al., 2012; Li et al., 2013; Patel et al., 2014, 2019). However, to date, these studies have focused on reflexive responses to unpredictable $f_o$ perturbations. When averaged across trials, these responses were highly reliable in the present study. It is possible that $f_o$ adaptive responses may have similar flexibility and may vary depending on uncontrolled factors such as internal states of the participant during an individual testing session.

Although the magnitude of $f_o$ adaptive responses was not reliable across testing sessions, categorical scoring of $f_o$ adaptive responses as typical versus atypical showed substantial test–retest agreement. This scoring method was implemented by Abur, Subaciute, Kapsner-Smith, et al. (2021) to detect differences in people with and without hyperfunctional voice disorders that were masked in comparisons of group means due to the presence of atypical responses at both ends of the distribution (i.e., compensating vs. following). Because there is substantial variability in $f_o$ adaptive responses in people with typical speech, scores that fall significantly outside of that distribution are remarkable. This scoring method has relevance for studying impaired auditory–motor control of speech in people with communication disorders. Studies with larger normative samples and/or meta-analysis of existing $f_o$ adaptation studies are needed to establish more precise cutoff scores for typical $f_o$ adaptive responses. Furthermore, normative data for $F_1$ reflexive responses could be used to assess the reliability of categorical measurement of articulatory feedback control using a similar approach.

One limitation of the present study is the difference in tasks (sustained phonation vs. words) for the $f_o$ versus $F_1$ experiments. This approach was used because most current studies of $f_o$ adaptive responses use sustained phonation, and most current studies of $F_1$ adaptive responses use speech or speechlike tasks. It is possible that participants’ $f_o$ feedforward targets may be less well defined for sustained phonation than their $F_1$ feedforward targets.
targets are for prolonged words. This could lead to more variability in $f_o$ responses than $F_1$ responses. However, the opposite pattern of reliability results observed for reflexive versus adaptive responses to $f_o$ and $F_1$ perturbations is not easily explained by task differences alone. Prior studies suggest differences in task may affect the magnitude of $f_o$ reflexive responses, with larger responses observed for singing than for speech (Natke et al., 2003), and for speech than for sustained phonation (S. H. Chen et al., 2007). These studies suggest $f_o$ feedback control is more active when the feedback target is more defined. In addition, responses to $f_o$ perturbations may vary depending on linguistic and musical experience, for example, singers (Zarate & Zatorre, 2008) and speakers of tonal languages (H. Liu, Wang, et al., 2010) in whom the $f_o$ auditory–motor control system is likely more developed. Future studies should examine the reliability of $f_o$ adaptation during different tasks and in participants with different linguistic and musical experience to see if responses are more stable depending on task and experience.

Reliability of assays of auditory–motor control of speech may be impacted by noise. $F_1$ and $f_o$ production is highly variable from trial to trial, and this variability may mask the stability of feedback and feedforward control mechanisms. It is possible that reliability may be improved by increasing the number of trials or other methodological changes that decrease the noisiness of the measures. Furthermore, the sample size in this study was relatively small, with $80\%$ power to detect medium (ICC $\geq .6$) effect sizes. A larger sample size could lead to different results.

In studies of auditory–motor control of speech, numerous methodological options exist for parameters including task and perturbation type, size, and timing. For example, the use of extended words in the $F_1$ tasks in this study means participants’ exposure to the perturbation was greater than in a natural word duration, which could have impacted the adaptive response. Outcomes such as the magnitude of responses may vary systematically with changes to task parameters, and it is conceivable that test–retest reliability could also be different depending on methodological choices. Future studies should examine the effects of such parameters on test–retest reliability.

Conclusions

Opposite patterns of reliability were observed for behavioral assays of feedforward and feedback control of voice ($f_o$) versus articulation ($F_1$). $f_o$ reflexive responses showed good reliability and $f_o$ adaptive responses showed poor reliability, whereas $F_1$ reflexive responses showed poor reliability and $F_1$ adaptive responses showed moderate reliability. However, a criterion-referenced categorical measurement of $f_o$ adaptive responses as typical versus atypical showed substantial test–retest agreement. Individual responses to some behavioral assays of auditory–motor control of speech should be interpreted with caution, especially when analyzing individual performance and/or change over time. Future studies should examine the reliability of criterion-referenced measures of $F_1$ adaptive responses as well as $f_o$ and $F_1$ reflexive responses. Furthermore, the opposite patterns of test–retest reliability observed for voice versus articulation add to growing evidence for differences in underlying neural control mechanisms.

Data Availability Statement

The data sets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grants R01 DC016270 (Cara E. Stepp), P50 DC015446 (Robert E. Hillman), T32 DC005361 (David J. Perkel), and F31 DC020359 (Mara R. Kapsner-Smith). The authors thank Chloe Thorburn, Callum Wilson, Aaron Moy, and Monique Tardif for assistance with data analysis.

References


Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the...


