JSLHR

Research Article

Kinematic Analysis of Speech Sound Sequencing Errors Induced by Delayed Auditory Feedback

Gabriel J. Cler,[a,b] Jackson C. Lee,[c] Talia Mittelman,[d] Cara E. Stepp,[a,b,d,e] and Jason W. Bohland[a,b,c]

Purpose: Delayed auditory feedback (DAF) causes speakers to become disfluent and make phonological errors. Methods for assessing the kinematics of speech errors are lacking, with most DAF studies relying on auditory perceptual analyses, which may be problematic, as errors judged to be categorical may actually represent blends of sounds or articulatory errors.
Method: Eight typical speakers produced nonsense syllable sequences under normal and DAF (200 ms). Lip and tongue kinematics were captured with electromagnetic articulography. Time-locked acoustic recordings were transcribed, and the kinematics of utterances with and without perceived errors were analyzed with existing and novel quantitative methods.
Results: New multivariate measures showed that for 5 participants, kinematic variability for productions perceived to be error free was significantly increased under delay; these results were validated by using the spatiotemporal index measure. Analysis of error trials revealed both typical productions of a nontarget syllable and productions with articulatory kinematics that incorporated aspects of both the target and the perceived utterance.
Conclusions: This study is among the first to characterize articulatory changes under DAF and provides evidence for different classes of speech errors, which may not be perceptually salient. New methods were developed that may aid visualization and analysis of large kinematic data sets.
Supplemental Material: https://doi.org/10.23641/asha.5103067

This special issue contains selected papers from the March 2016 Conference on Motor Speech held in Newport Beach, CA.

Fluent speech incorporates the reception of speakers' own productions via auditory feedback. When external auditory feedback is delayed by approximately 200 ms, speakers reduce their speech rates and produce errors in speech output, including disfluencies and phonological sequencing errors (Fairbanks, 1955; Yates, 1963). However, the delayed auditory feedback (DAF) effects remain poorly understood and cannot be readily explained by contemporary models of speech motor control (e.g., Guenther, Ghosh, & Tourville, 2006; Hickok, 2012; Hickok, Houde, & Rong, 2011; Houde & Nagarajan, 2011; Saltzman & Munhall, 1989; Tilsen, 2013). These models tend to lack either the specification of mechanisms for sequencing multiple sounds in a speech plan (Guenther et al., 2006; Hickok, 2012; Hickok et al., 2011; Houde & Nagarajan, 2011), which are essential to explaining the observed discrete serial order errors, or do not explicitly address the use of online auditory feedback (Saltzman & Munhall, 1989; Tilsen, 2013). Although the Gradient Order DIVA model (Directions Into Velocities of Articulators; Bohland, Bullock, & Guenther, 2010) extends the DIVA speech motor control framework (Guenther et al., 2006) to address how the brain may plan and produce sequences of speech sounds, it does not yet account for the effects of DAF due to an incomplete treatment of the auditory-perceptual system.

Speech Sequencing and Auditory Feedback: Theoretical Framework

The production of speech sequences is thought to involve a phonological encoding stage, in which the content of the planned utterance is represented at an abstract level (i.e., a phoneme sequence), which is then used to address and select learned motor programs for articulation

[a]Graduate Program for Neuroscience–Computational Neuroscience, Boston University, MA
[b]Department of Speech, Language, and Hearing Sciences, Boston University, MA
[c]Department of Health Sciences, Boston University, MA
[d]Department of Biomedical Engineering, Boston University, MA
[e]School of Medicine, Department of Otolaryngology–Head and Neck Surgery, Boston University, MA
Correspondence to Gabriel J. Cler: mcler@bu.edu

Editor: Yana Yunusova
Associate Editor: Jeffrey Berry
Received June 14, 2016
Revision received October 7, 2016
Accepted November 16, 2016
https://doi.org/10.1044/2017_JSLHR-S-16-0234

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

(e.g., Levelt, Roelofs, & Meyer, 1999). Under this view, coordinated neural mechanisms must sequentially select the appropriate phonological units from a planning buffer, while concurrently activating lower-level sensory-motor programs that drive the production of individual syllables (Bohland et al., 2010). At this lower level, which is described computationally by the DIVA model, continuous external auditory feedback is compared against stored sensory expectations specified as formant trajectories over time. It is understood, however, that external auditory feedback is also simultaneously used to monitor speech output for errors at multiple linguistic levels, including for the detection and correction of discrete segmental and suprasegmental errors (Levelt, 1983; Postma, 2000). A conservative estimate is that healthy adults make errors in the serial sequencing of speech sounds at a rate of approximately one to two per 1,000 words spoken (Garnham, Shillcock, Brown, Mill, & Cutler, 1981; Hotopf, 1983), and they perform online corrections of perhaps 50% of naturally occurring speech errors (Nooteboom, 1980). Note that such errors are theoretically distinct from subphonemic mismatch errors (i.e., distorted productions), which are characterized by a graded rather than categorical deviation from the target sound.

Under DAF, a relatively large fraction of the evoked speech errors appear, at least perceptually, to be discrete and categorical, suggesting that the mechanism leading to error may be distinct from the feedback controller suggested by DIVA and other related models to steer individual productions toward the intended sensory consequences. Instead, such DAF-induced errors may highlight a higher level auditory feedback loop in which the incoming sound sequence is compared against the planned sequence at a more abstract level; any detected errors at this level could then be used to invoke activity changes within the (phonological) speech-planning buffer, for instance, to correct a naturally occurring error. Because of the aberrant timing of feedback introduced under DAF, categorical mismatches at the level of the sound sequence will be frequent and may lead to a reset, perseveration, or misordering of elements in the speech plan, thereby driving production anomalies that are often perceived as discrete sequencing errors.

Experimental Manipulations of Auditory Feedback

During speech, an auditory signal is transmitted to both the intended listener and back to the speaker as auditory feedback. This feedback is critical for learning and maintaining sensorimotor mappings needed for effective speaking, as is evidenced by early childhood deafness impairing the typical acquisition of speech (Oller & Eilers, 1988) and deleterious effects in adults who lose hearing after acquiring language (Perkell et al., 2001). A wealth of experimental evidence suggests that altered auditory feedback has direct effects on relatively low-level parameters of the speech motor controller, driving changes in vocal intensity, reductions in speaking rate, and changes in phonemic contrasts under masking noise (Lane & Tranel, 1971; Perkell et al., 2007; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988), as well as compensatory responses to shifts in formant frequencies or fundamental frequency (Cai, Ghosh, Guenther, & Perkell, 2011; Purcell & Munhall, 2006; Tourville, Reilly, & Guenther, 2008; Villacorta, Perkell, & Guenther, 2007; Xu, Larson, Bauer, & Hain, 2004).

When auditory feedback is altered by inserting an artificial delay (typically of approximately 200 ms), a large number of speech errors emerge, including both distorted productions and, as noted previously, discrete sequencing errors. The extent of these DAF-induced speaking effects appears to be highly variable across individuals (Burke, 1975; Chon, Kraft, Zhang, Loucks, & Ambrose, 2013). Previous work using a highly controlled experimental protocol that mirrors the one used here shows, for example, that while DAF results in serial speech errors in more than 60% of trials in some participants, others are more robust to the manipulation, with such errors occurring in only approximately 10% of trials (Malloy, Nistal, & Bohland, 2014). Malloy et al. (2014) found a selective increase in phonological errors involving vowels or whole syllables with increasing delay, whereas consonant errors (the most common naturally occurring error unit) were unaffected. Discrete errors in which whole phonemes or syllables are inaccurately sequenced have also been frequently reported in past studies using DAF (Chon et al., 2013; Fairbanks & Guttman, 1958; Yates, 1963).

Speech Error Identification and Electromagnetic Articulography

The literature surrounding naturally occurring and/or laboratory-induced slips of the tongue has historically converged on the segment as the basic unit represented in the phonological speech plan (Dell, 1986; Nooteboom, 1973; Shattuck-Hufnagel, 1983; Shattuck-Hufnagel & Klatt, 1979). This is primarily due to the perceived well-formedness of most speech errors. The evaluation of speech errors from audio recordings, however, may be influenced by listeners' perceptual biases, as well as by the conventions used for phonetic transcription (Cutler, 1981; Frisch & Wright, 2002; Pouplier & Goldstein, 2005). Previous work using electromagnetic articulography (EMA), with normal auditory feedback, has suggested that many errors perceived to be categorical may involve subphonemic articulatory errors, including the coproduction of multiple segments (Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; Pouplier, 2007; Pouplier & Hardcastle, 2005). Thus, it is important to determine if errors that are induced by DAF and judged by listeners to be categorical represent *pure* sequencing errors or involve subphonemic alterations to the articulatory output because the two classes of errors may arise from different mechanisms, as noted previously. Understanding the nature of DAF-induced changes to articulation will inform modeling treatments that address the multiple levels of interactions between the perceptual and production systems in running speech.

The only previous study, to our knowledge, of speech kinematics under DAF used infrared light-emitting diodes

placed on the lips and jaw only (Sasisekaran, 2012). In this study, participants read nonwords with typical auditory feedback, under delay of 200 ms, or with gated feedback in which the auditory feedback was turned off and on at 2 Hz. Lip aperture was evaluated in the different conditions, and aperture variability was higher under delay than in nondelay or gated conditions. Note that only trials without perceptual errors were evaluated, and the participants with the highest error rates were excluded. Indeed, most existing articulographic studies employ measures of the kinematics of speech (e.g., spatiotemporal index [STI]) that disregard speech error data, with a few notable exceptions (Goldstein et al., 2007; Pouplier, 2007); as such, quantitative methods for conducting kinematic analyses of errorful speech are currently lacking.

The primary goals of this study were to develop and apply data-driven methods to determine how DAF affects syllable articulations in a highly controlled speaking task and to determine how articulations varied during productions deemed by listeners to be correct productions, distorted versions of correct syllables, or syllables that involved categorical sound errors. Though limited in the number of participants (eight), this approach produced hundreds of speech tokens per participant, enabling meaningful within-participant analyses. Listeners transcribed the speech produced, and syllables were classified as correct productions, graded distortions, or categorical errors. We examined nonerror trials using STI and used a multivariate approach to analyze both nonerror and errorful speech tokens. We examined the "landscape" of these productions reflecting trial-to-trial articulatory variability and compared productions perceived to be error free with and without DAF. We also examined the articulatory kinematics of distortion and categorical errors in relation to canonical productions of the target and transcribed syllables. Although aspects of this work are exploratory in nature and will require further study, here we have demonstrated a novel approach to assessing errors under DAF, while providing results that demonstrate a mixture of both graded and strictly categorical speech errors.

Method

Participants

Eight healthy young adults (three women, five men; *M* age = 24.5 years, range 19–33 years), participated in the experiment. All participants were native American English speakers and reported no history of speech, language, or hearing impairments. Participants provided written consent in compliance with the Boston University Institutional Review Board.

Data Collection

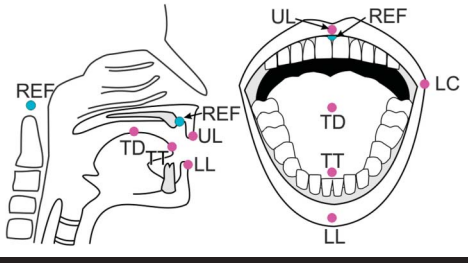
Simultaneous speech acoustics and articulator position data were recorded by using the NDI Wave Speech Research System (Northern Digital Inc., Waterloo, Ontario, Canada) in a sound-attenuating booth. Acoustic data were sampled at 22 kHz, and kinematic data were sampled at 100 Hz. Eight EMA sensors were used, each of which captured 5 *df* measurements. Three sensors used for head correction were placed on the gingiva of the upper incisors and the left and right mastoid processes. The position of each participant's maxillary occlusal plane with respect to the reference sensors was recorded by using a plastic mouth guard with three sensors attached: one at the front center directly beneath the diastema of the front teeth and two placed symmetrically on either side, in the back of the mouth guard, to fit beneath the molars. After the maxillary occlusal plane was measured, the mouth guard was removed, and five EMA sensors were attached to articulators: tongue tip (TT), tongue dorsum (TD), upper lip (UL), lower lip (LL), and left corner of the mouth (see Figure 1). The sensor locations were chosen a priori to maximally differentiate between the six phonemes (/a/, /i/, /u/ and /b/, /d/, /z/; see Experimental Protocol section for more details). The TT was placed approximately 5 mm from the actual anterior tip of tongue. The TD was placed as far back as was feasible, with individual variation, but at least 2 cm from TT.

The reference sensor on the gingiva and the five articulatory sensors were attached with dental adhesive (high-viscosity PeriAcryl, GluStitch Inc., Delta, BC, Canada), whereas the two sensors on the mastoid processes were attached with double-sided tape and medical tape. After sensors were attached, participants were fitted with a head-worn condenser microphone (Shure WH30XLR, Shure Incorporated, Niles, IL), electrostatic insert earphones (Sensimetrics S14, Sensimetrics Corporation, Malden, MA), and dielectric earmuffs (Howard Leight Thunder T3, Honeywell Safety Products, Smithfield, RI) that had a noise reduction rating of 30 dB.

Experimental Protocol

Acoustic signals were transmitted from the head-mounted microphone to an external sound device (M-Audio Fast Track Ultra, M-Audio, Cumberland, RI) connected to a laptop computer. The signal was amplified by +8 dB to attempt to overcome bone conduction (Cornelisse, Gagne, & Seewald, 1991) and transmitted to the earphones after being temporally modified by an experimentally specified delay. Auditory feedback delays were implemented

Figure 1. Positions of electromagnetic articulography (EMA) sensors. Reference (REF) sensors were placed on left and right mastoids and gingiva of upper incisors. Articulatory sensors were placed on the tongue tip (TT), tongue dorsum (TD), upper lip (UL), lower lip (LL), and left corner of the mouth (LC).



using PsychPortAudio (Brainard, 1997), a software sound interface available for the MATLAB Psychophysics Toolbox (MathWorks, Natick, MA), which utilizes Audio Stream Input/Output drivers to obtain high temporal precision and low latency playback. The feedback provided to the participant in each trial was either presented in near real time (sound processors introduced an approximately 14-ms delay; 25% of trials) or delayed by 200 ms (pseudorandom selection of 75% of the trials).

Participants repeated nonsense syllable sequences at a constrained pace. In each of the trials, participants repeatedly produced one of 12 pseudorandomly chosen $C_1V_1C_2V_2C_3V_3$ nonsense sequences, with vowels chosen from /a/, /i/, and /u/ and voiced consonants chosen from /b/, /d/, and /z/. No phonemes were repeated in a given sequence. The 12 sequences were a priori chosen from the possible set to constrain productions to six consonant–vowel (CV) syllables and 12 vowel–consonant–vowel (VCV) transitions to produce as many repetitions of each utterance as possible; the stimuli sequences were /biduza/, /bizadu/, /budazi/, /buzida/, /dabuzi/, /dazibu/, /dubiza/, /duzabi/, /zabidu/, /zadubi/, /zibuda/, and /zidabu/. Stimuli were orthographically displayed on a monitor (e.g., "dah boo zee") for 2 s before being removed and replaced with a visual metronome. One of the most commonly observed effects of DAF is a near immediate reduction in speech rate, a potential compensatory strategy (Black, 1951) that speakers use to reduce error occurrences (i.e., trading off speed for accuracy). Such rate reductions, however, may obscure the direct effect DAF has on serial speech at typical speaking rates. The visual metronome was, therefore, used to cue participants to produce syllables at a steady rate in an attempt to counteract this compensatory strategy. This cue appeared as three circles positioned horizontally on the screen, sequentially changing colors at 5 Hz, and remaining on the screen for the duration of a 3-s production period. During this production period, participants repeatedly produced the CVCVCV sequence with or without DAF.

Participants completed 360 trials across six runs, in between each of which was approximately a 3- to 5-min rest period. Three participants had sensors detach during the experiment; in these cases, the run was halted, and the sensor was reattached. During analysis, the larger set of data (i.e., before or after reattachment), was retained for analysis and the remaining data were excluded. Reattachment of the sensor, even when placed within millimeters of its original position, generally introduced significant artifacts during analysis.

Data Analysis

Transcription

Authors used the speech analysis software Praat (Boersma & Weenink, 1996) to transcribe and mark onsets of syllables, the primary unit of transcription and analysis here, within the recorded audio for each trial by viewing the waveform and spectrogram and listening to the sample. Syllable onsets were located by marking the onset of high energy in the frequency spectrum, due to the burst of /d/ and /b/ or the onset of the frication of /z/. Listeners then transcribed each syllable heard with a closed set of phonetic symbols in which productions of the set of phonemes (/b/, /d/, /z/, /a/, /i/, /u/) that were considered typical were marked. Any atypical productions (of either consonants or vowel segments) were marked with a separate symbol (@). Custom MATLAB software was used to automatically compare the transcription to the stimulus presented to the participant to identify and classify errors.

Auditory-Perceptual Analysis and Automated Error Identification

The second author transcribed recordings for all trials (consisting of three to four repetitions of $C_1V_1C_2V_2C_3V_3$ sequences) and participants, blinded to the stimulus. Any trials in which transcriptions matched the stimulus exactly (i.e., the trial contained multiple correct productions of the entire sequence only) were marked (using automated scripts as noted previously) as nonerrors. Trials in which the transcription differed from the stimulus within the first three syllables were defined as *misremember* errors and discarded; these discrepancies could be caused by the delay (and indeed their frequency of occurrence did increase under DAF) or by the participant forgetting which sequence they had been prompted to produce.

All remaining trials in which the first transcription differed from the stimulus, suggesting an error occurred, were transcribed by the third author. Trials in which the first and second transcriptions differed in the identified type of error were transcribed by the first author. Any error trials that did not then have agreement between two of the three transcribers were discarded. Any transcribed syllables before the occurrence of an error syllable were considered nonerrors, and all syllables in the trial after the first transcribed error syllable were discarded. All transcribers were native English speakers.

Errors were automatically detected and denoted as containing within-set phonemes or out-of-set phonemes. All syllables that were marked with an out-of-set character (@) were separately recategorized by the second transcriber into two groups: (a) syllables containing an ill-formed in-set phoneme (i.e., a distortion) or (b) syllables containing a well-formed out-of-set phoneme. All error productions were categorized as in Table 1, including categorical errors, distortions, out-of-set errors, and artificial stutters. Two main classes of errors were selected for further analysis. The first were considered *categorical* errors, in which transcribers perceived the syllable as a properly formed CV with phonemes from the stimulus set, but these perceived phonemes did not match the target syllable within the prescribed stimulus. The second class of errors were distortions, in which either the consonant or the vowel was perceptibly distorted but did not cross perceptual categories to be perceived as a different vowel. Well-formed versions of out-of-set phonemes (e.g., d^) were not examined further, as there were no nonerror data for such syllables.

Table 1. Types of discrepancies between stimulus and transcription.

Error type	Description	Example stimulus: bi du za
Categorical	A produced syllable contains two phonemes from the stimuli set (/b/, /d/, /z/, /a/, /i/, /u/) but is different from the stimulus syllable. Can include vowel, consonant, or full syllable repetitions, anticipations, or exchanges	bi **b**u za bi d**a** za
Distortion	A syllable contains a phoneme that is not clearly identifiable	bi d**@** za
Out of set[a]	A syllable contains a phoneme that is perceptibly well-formed but was not present in the stimulus set	bi du **z^**
Artificial stutter[a]	A disfluency where two consonants are produced without a clear vowel production	bi **d-du** za
No agreement[a]	No two out of three transcriptions agreed on the nature of the error	
Misremember[b]	An error in the first three syllables of a trial; could be caused by DAF or by the participant forgetting which sequence to produce	

Note. Categories of errors detected and analyzed under DAF in this study, with an example errorful sequence of each type.
[a]This type of error was not analyzed further in this study. [b]Trials classified as misremembers were discarded and not counted as errors or nonerrors.

Analysis of Syllable Durations

The duration for each syllable was calculated from the marked onset (see the data analysis section of Method for more details) to the onset of the subsequent syllable. To determine whether average syllable duration was affected by experimental factors, durations of nonerror syllables were analyzed by using an analysis of variance with main factors of delay (0 or 200 ms), syllable (/bi/, /bu/, /da/, /du/, /za/, /zi/), syllable order in the stimulus sequence (one to three), and Participants 1–8 (random factor), with all possible interactions.

EMA Data Preprocessing

EMA data were exported from the NDI WaveFront software and imported into MATLAB. Data were low-pass filtered with a third-order Butterworth filter with a 5-Hz cutoff for the reference sensors and 20-Hz cutoff for the articulatory sensors, following Tiede et al. (2010). To correct for head motion, the kinematic data were rereferenced to each individual's articulatory space, with the origin between the back molars and behind the diastema of the upper central incisors.

For STI. EMA data corresponding to each of the 12 sequence types (e.g., /dazibu/) were extracted beginning at the first marked syllable onset and ending after the final syllable offset, which was also the onset of the following sequence. Sequences were amplitude normalized and then time normalized to 1,000 data points per sequence via spline interpolation (Smith, Johnson, McGillem, & Goffman, 2000).

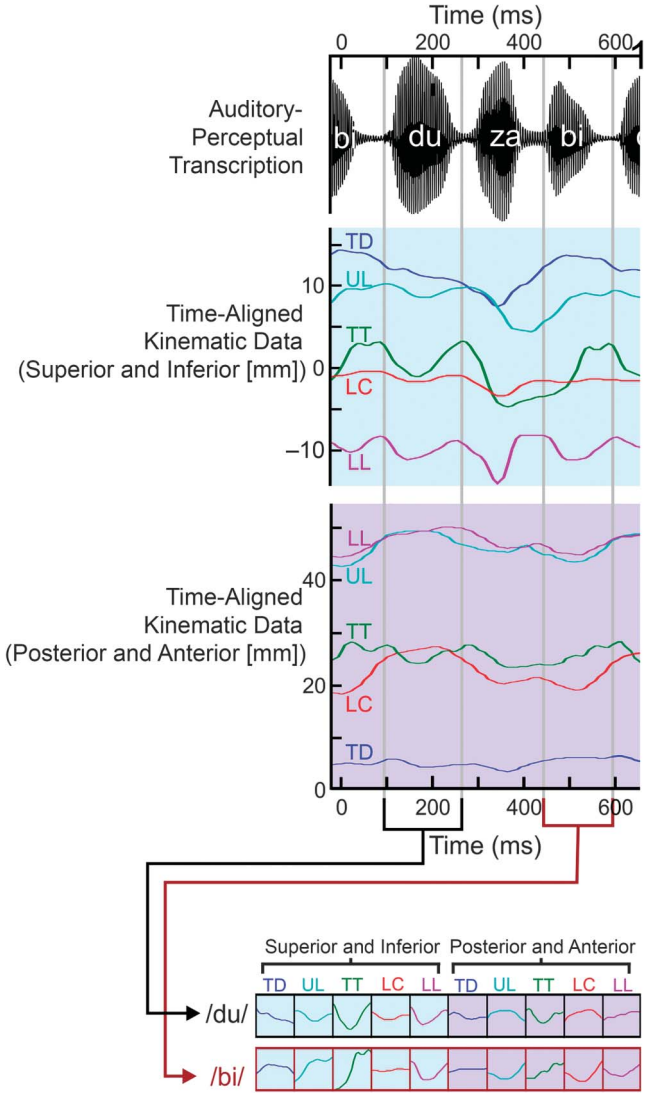
For all other measures. EMA data corresponding to individual syllable productions were extracted starting at the marked syllable onset and ending at the subsequent syllable onset. To accommodate different syllable durations, the EMA data for each syllable were then time normalized by resampling and linearly interpolating data such that each was 100 time samples long. Syllable boundaries were then shifted to incorporate 30% of the previous syllable and 70% of the current syllable to ensure that all articulatory movements related to the onset consonant were included in the sample;[1] thus, the first syllable of each trial was also discarded. Syllables transcribed as correct productions were discarded if they were longer in duration than 4 *SD*s above the mean (for that participant) or if they had kinematic excursions greater than 100 mm within the syllable in any sensor or had missing data, indicating a measurement error by the EMA device. Although kinematic data were captured in three dimensions (posterior and anterior, left and right, and inferior and superior), qualitative comparisons of analyses with and without the left and right dimension indicated that little information was provided by this dimension, and thus all further analyses were completed with measurement data from only the posterior and anterior and inferior and superior dimensions.[2] EMA data from syllables were converted (through simple concatenation) from a matrix of Time (100 samples) × Space (posterior-anterior and inferior-superior) × Sensor (TT, TD, UL, LL, LC) into one high-dimensional feature vector in which each feature (element in the vector) corresponded to the data from one sensor at one time point from one spatial dimension. Figure 2 shows an example of the data used, including the transcription time-aligned to kinematic and acoustic data, and illustrates how the kinematic data were reordered into a feature vector.

Kinematic Analyses

Analyses included calculating the difference in STI between nonerror sequences with and without delay, calculations of distance between syllables in high-dimensional space, visualizing the variability of articulation in nonerror and error syllables, and classifying errors into different

[1]The average syllable duration under no delay was 203 ms and under delay was 235 ms. Using 30% of the previous syllable incorporates hold periods consisting of 60–70 ms, which is well within the range of typical hold periods (Hixon, Weismer, & Hoit, 2008).
[2]This was surprising given that the corner of the mouth sensor location was chosen specifically to capture movement in the left and right dimension between /i/ and the other vowels.

Figure 2. Data preprocessing methods (trial segment from Participant 4 [P4]). Top panel shows acoustic signal overlaid with transcription, with grey vertical lines indicating where the transcriber marked the syllable onsets. Second and third panels show kinematic data time-aligned to acoustic signal (second panel: five articulatory sensors in superior and inferior dimension; third panel: five articulatory sensors in posterior and anterior dimension). The bottom panel shows how kinematic data are extracted on the basis of marked onset boundaries, time normalized to 100 points per syllable, and restructured into one vector per syllable in which each of the 1,000 features represents one sensor's position in one dimension at a particular time in the syllable. TD = tongue dorsum; UL = upper lip; TT = tongue tip; LC = left corner of the mouth; LL = lower lip.



error profiles. All analyses were performed within-participant only.

Differences in spatiotemporal index. Sequences under delay and not under delay were evaluated separately within utterance type, sensor, and dimension. STI was calculated as the sum of standard deviations at 2% intervals over sequences time normalized to 1,000 data points. As there were different numbers of sequences in the delay and no-delay conditions, the larger category was bootstrapped. If there were, for example, 30 /dazibu/ sequences under delay and 20 /dazibu/ sequences without delay, one STI value was calculated over the smaller category (in this case, without delay) for each dimension of each sensor. The STI for the larger category was calculated as the mean of 100 bootstrapped STI values. That is, random sets of sequences (equal to the size of the smaller category) were drawn from the larger category; STI was calculated over each set and averaged. Within utterance, sensor, and dimension, the STI of sequences produced with no delay were subtracted from the STI of those produced under DAF. The resultant differences in STI (ΔSTIs) were averaged over utterance type and then over dimension and sensor to provide an overall measure of the difference in articulatory variability between delay and nondelay trials for each participant.

Distance calculations. A global measure of distance was used to characterize the differences between individual syllable productions. The distance between any two syllable feature vectors represents how different the kinematic traces are by comparing a value representing a single time, dimension, and sensor to another trace's value at that same time, dimension, and sensor. The larger the distance between two traces, the more different they are. Distance measures were calculated with the Manhattan (L_1-norm) distance formula. Thus, the distance between two syllable productions (e.g., $\vec{x}$, $\vec{y}$), each represented by a vector with N = 1,000 elements is given by

$$d_1\left(\vec{x},\vec{y}\right) = \sum_{i=1}^{N} \left|\vec{x}_i - \vec{y}_i\right|.$$

For each participant, a mean nonerror feature vector representation was calculated across all nonerror productions for each of the six syllable classes. These class centroids were used as a representation of a prototypical production. Distances were calculated from each individual syllable vector to the mean feature vector (centroid) corresponding to its perceived syllable class. One-tailed, two-sample Kolmogorov–Smirnov tests were completed to compare the distributions of these distances for syllables produced with and without DAF; these tested the hypothesis that the mean and/or variance of the distance from centroid (i.e., articulatory variability) was larger for syllables produced under delay than those with no delay.

Visualization of articulatory variability. The distances between all pairs of syllable feature vectors were calculated, and nonmetric multidimensional scaling (MDS; Kruskal, 1964) was used to visualize these relationships in two dimensions. In these plots, the distance between data points (each representing one syllable production) in two-dimensional space is monotonically related to the distance between all data points in the high-dimensional feature space (see Distance calculations section). Data were plotted without reference to the transcribed syllable identity, but the color and marker for each production was indicated post hoc to reflect their perceived class. Thus, any groupings

of syllables are due purely to their differences and distances in high-dimensional space, rather than to any explicit data clustering process.

Classification of error profiles. To broadly categorize error productions into a set of *error profiles* (e.g., alike only the stimulus class, alike only the produced class, alike both, or other), we performed classification of the kinematic data by using machine learning tools. For each of the six canonical syllable classes in each participant, one quadratic discriminant analysis classifier was trained to discriminate nonerror syllables labeled as *in class* (target syllable) from those labeled as *out of class* (remaining five syllables). To test reliability of the classifiers on nonerror productions, fivefold cross-validation was performed on each syllable class and then averaged across syllables within participant. Six quadratic discriminant analysis classifiers were then retrained by using a participant's entire set of nonerror syllables, as noted previously. Each syllable production transcribed as a categorical error was then classified by each of the six classifiers as in class or out of class and assigned to an error profile. For example, if an error was classified as in class of its stimulus syllable and out of class of the remaining five syllables, it was assigned the error profile *alike only stimulus*. If it was in class for both the stimulus syllable and perceived syllable and out of class for the remaining four syllables, it was assigned *alike stimulus and perceived.* Distortion errors were similarly classified and assigned to error profiles. Error profile types were then tallied across participants to determine the overall distribution of these profiles in the entire data set.

Results

Auditory-Perceptual Analysis

Participants completed 240–360 trials. After preprocessing data as described previously, participants had between 2,453 and 3,948 (M = 2939, SD = 615) syllables in which transcriptions matched the target stimuli (nonerrors) and between 11 and 133 (M = 85, SD = 47) syllables marked as errors. Each participant had an additional four to 70 trials discarded as misremember errors (225 trials total, with nine to 12 syllables each), as well as one to 21 syllables (52 total) in which two of three transcribers could not agree and were thus not analyzed further. We note that although our design required participants to produce sequences from memory and thus required discarding some trials due to these ambiguous misremember errors, it eliminates uncontrolled variables related to how participants use visual inputs when stimuli can be directly read. Further, the effects of DAF have been shown to differ under reading and conversational speech conditions (e.g., Corey and Cuddapah, 2008), so to focus our analysis exclusively on speech motor mechanisms, we opted to remove the visual aspect of this task during the experimental manipulation.

The overall number of errors and the distribution of the types of errors produced varied between participants, as has been previously reported (Malloy et al., 2014).

Figure 3. Distribution of errors with and without DAF for the eight participants. Left column for Participants 1–8 (P1–P8) shows the error rates for each of the error types with no delay, and the right column shows the error rates under 200-ms delayed feedback.

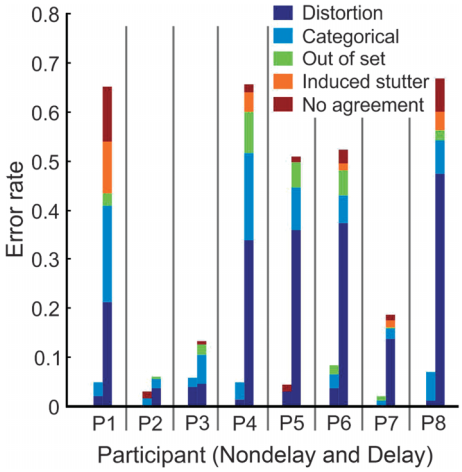


Figure 3 shows the distribution of errors among participants. Error rates and percentage increase in error rate under delay are summarized in the second column of Table 2; note that while raw error rate is lowest in Participants 2, 3, and 7 (P2, P3, and P7), the percentage increase in error rate under delay was lowest in P2, P3, and Participant 6 (P6), suggesting that these participants were least affected by delay. Although the visual metronome was intended to help participants maintain a consistent syllable rate, an analysis of variance showed main effects of delay (0 or 200 ms), syllable (/bi/, /bu/, /da/, /du/, /za/, /zi/), and syllable order in the sequence (one to three) on syllable duration, with multiple significant interactions (see Table 3). The mean duration of syllables with no delay was 203 ms (SD = 46 ms),

Table 2. Statistical differences in delay and nondelay kinematic variability.

Participant	Error rate under delay; percentage increase from nondelay	KS test statistic
1	0.65; 1,230	.23**
2	0.06; 115	.03
3	0.13; 130	.00
4	0.66; 1,278	.25**
5	0.51; 1,106	.16**
6	0.52; 522	.05
7	0.19; 867	.05*
8	0.71; 937	.23**

Note. Error rates under delay and percentage increase in error rate under delay for each participant. The right-most column shows results of one-tailed, two-sample Kolmogorov–Smirnov (KS) tests that compared the distributions of distances from the category centroid for syllables produced with and without DAF.

*p < .05 and **p < .001 indicate that the kinematics of delay trials are significantly farther from the mean and/or have more variability from the mean than nondelay trials.

Table 3. ANOVA results for syllable duration.

Effect	*df*	Adjusted sum of squares	η_p^2	*F*
Delay	1	5.79	0.19	18.12*
Order	2	10.87	0.30	21.79**
Participant	7	3.65	0.13	0.93
Syllable	5	18.79	0.42	88.60**
Delay × Order	2	0.78	0.03	7.68*
Delay × Participant	7	2.30	0.08	5.24*
Delay × Syllable	5	0.47	0.02	6.68**
Order × Participant	14	3.60	0.12	4.74*
Order × Syllable	10	0.14	0.01	2.70*
Participant × Syllable	35	1.53	0.06	2.69*
Delay × Order × Participant	14	0.73	0.03	14.95**
Delay × Order × Syllable	10	0.08	0.00	2.25*
Delay × Participant × Syllable	35	0.51	0.02	4.16**
Order × Participant × Syllable	70	0.37	0.01	1.52*
Delay × Order × Participant × Syllable	70	0.25	0.01	2.74**
Total	19,772			

Note. Durations of nonerror syllables were analyzed by using an analysis of variance (ANOVA) with main factors of delay (0 or 200 ms), syllable (/bi/, /bu/, /da/, /du/, /za/, /zi/), syllable order in the stimulus sequence (one to three), and Participants 1–8 (random factor), with all possible interactions. Shown are the degrees of freedom, adjusted sum of squares error, partial eta-squared value, and *F* value for each effect.

*$p < .05$. **$p < .001$.

while the mean duration under DAF was 235 ms (*SD* = 72 ms). This change in production rate under delay is approximately 0.7 syllables/s; this is smaller than changes observed in studies without a metronome. For example, Stuart, Kalinowski, Rastatter, and Lynch (2002) found that participants' speaking rate slowed down at 200-ms delay by approximately 1.2 syllables/s in a normal speaking pace and approximately 1.3 syllables/s in a fast speaking pace.

Kinematic Analyses

Distribution of Nonerror Productions

Figure 4 shows nonmetric MDS representations of the landscape of syllable productions for three example Participants 8 (P8), 4 (P4), and P2 (all participants included in Supplemental Material S1). These plots show all nonerror productions for each participant, and the distances between the markers in this two-dimensional plot are monotonically related to the dissimilarity of syllable productions in the high-dimensional feature space. Note that the degree of clustering by syllable differs across participants, with some clusters overlapping more than others. Clusters with common vowels are often nearest, although some participants' data resulted in clusters in which the consonant was more dominant in determining cluster overlap. Figures 4A and 4B show two participants in which clusters are largely segregated by syllable class. Figure 4C shows a participant in which clusters overlap somewhat in the two-dimensional representation and in which the overlap is primarily governed by vowel; for example, note that clusters for the syllables /za/ and /da/ overlap slightly, as do clusters for /zi/ and /bi/.

In addition to color coding by syllable, color intensity is used to denote whether the syllable was produced with DAF (lighter) or without (darker). In Figures 4A and Figure 4B, note that the syllables produced under DAF appear to show much more variation from their class centroids than the nondelay syllables. This suggests that there may be more articulatory variability in the productions under delay even in utterances that are perceived as error free. The participant in Figure 4C, however, shows no apparent difference in variability under delay and nondelay. The participant in Figure 4C was one who did not show large differences in error production under delay (see Figure 3). Statistical testing shows that five out of eight participants had productions under DAF that had larger and/or more variable distances from the mean than productions without DAF (see Table 2). It is interesting to note that the three participants who did not show statistically significant differences in these distributions (P2, P3, and P6) were those with the smallest percentage increase in error rate under delay.

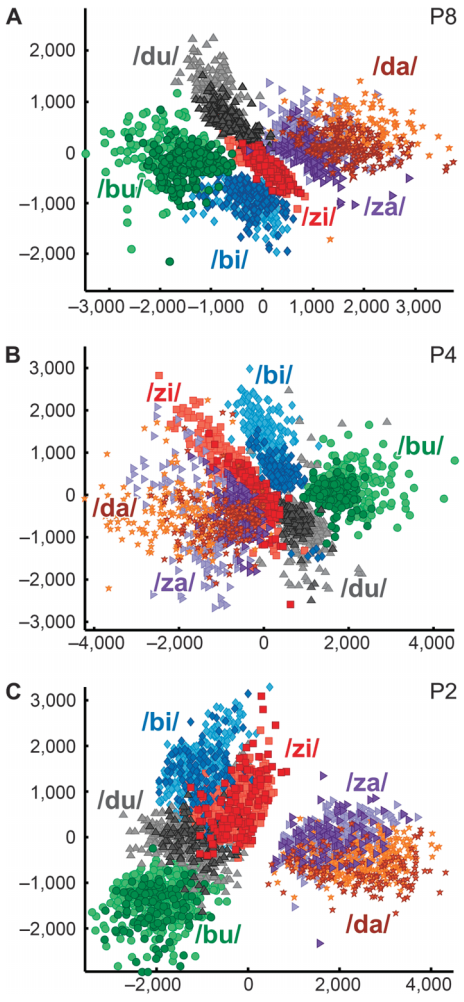
STI

STI was also calculated for CVCVCV sequences to compare variability in productions made with and without DAF and for comparison with our current approach. Figure 5 shows the difference in STI between delay and nondelay sequences (differences calculated within utterance type, sensor, and dimension before being averaged). Note that the participants with higher error rate increases under delay also tend to have higher STIs under DAF, while those participants with lower error rate increases have similar STIs regardless of delay. Four of the five participants who had a significant increase in articulatory variability (see Table 2) also had an increased STI under DAF.

Distribution of Error Productions

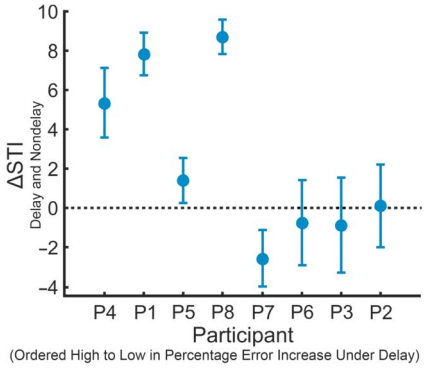
Figure 6 shows syllables transcribed as errors plotted on nonerror syllable clusters for Participant 1 (P1). Figure 6A shows all distortions color coded by the target

Figure 4. Example multidimensional scaling (MDS) plots of nonerror productions for Participants 8, 4, and 2 (P8, P4, and P2). Each data point indicates one nonerror syllable production, color coded by syllable class. Distance between points is monotonically related to the L1-norm distance between high-dimensional kinematic feature vectors recorded and resampled for each syllable. Blue diamonds: /bi/, green circles: /bu/, orange stars: /da/, black triangles: /du/, purple arrows: /za/, red squares: /zi/. Lighter markers indicate trials produced under delay, while dark markers show trials with no delay. MDS plots for the remaining participants are included in Supplemental Material S1.



stimulus. For example, all nonerror /bu/ syllables are plotted in green, and any syllables that were supposed to be /bu/ but were transcribed as a vowel distortion (/b@/) or consonant distortion (/@u/) are marked with green-crossed circles. Note that these were in many cases proximal to the nonerror stimulus syllable cluster, suggesting some overall resemblance to the target syllable. Figure 6B shows the same participant's categorical errors that matched one of the six CV syllables used in the experiment. Here, the color of the circle indicates the stimulus for that syllable, while the color of the center "x" indicates the syllable that was perceived. Three specific examples of categorical errors are highlighted here: one in which the error is near the center of the perceived cluster /bi/ (blue-dashed box), one in which the kinematics are near the boundary between the perceived and stimulus syllable clusters (red solid box), and one in which the kinematics are closer to the stimulus than to the perceived syllable cluster (black-dotted box).

Figure 5. Differences in spatiotemporal index (ΔSTI) by participant. Participants 1–8 (P1–P8) are ordered by percentage increase of error rate under delay (Table 2). ΔSTIs represent the difference in STI between delay and nondelay sequences (differences calculated within utterance type, dimension, and sensor before being averaged). Error bars represent standard error of the mean difference of STI calculated over the sensors.



Proximity of Distorted Productions to All Syllable Classes

Figure 7 shows spider plots that compare several example distortion errors with canonical (centroid) productions of all six syllable classes. For these plots, centroids of each syllable class were calculated (as in Distance calculations section). For each syllable production plotted here, the distances between that syllable's feature vector and each of the six centroids were found. The inverse of distance (dissimilarity) was used as a proxy for similarity. This similarity measure suggests that the closer to the end of each axis, the more similar (smaller distance) a production is to the mean nonerror production of that class. The midpoint (origin) of each axis is a similarity of 0 (infinite distance from the centroid), and the end point is set to the highest similarity of any nonerror production to the centroid of that syllable class. Figure 7A shows three distortions of the stimulus syllable /zi/ (target icon) from P8 plotted with a randomly chosen set of nonerror productions of this same class from the same participant. In two of these examples, the distorted syllable productions were clearly more similar to nonerror /zi/ productions than to any other syllable, while the third showed reduced similarity to any of the six canonical syllables. Figure 7B shows distortions of /bi/ made by P4 with three different patterns: one is similar to canonical /bi/ productions, one is more similar to a canonical /bu/, and one is relatively equidistant from all clusters and thus not like any of the prototypical productions.

Proximity of Categorical Error productions to All Syllable Classes

Figure 8 shows three examples of syllables transcribed as categorical errors with different relationships to

Figure 6. Relative location of error productions for Participant 1 (P1). (A) shows errors perceived to be distortions along with all nonerror syllables; the color of the distortion marker indicates the stimulus syllable for that production. (B) shows categorical errors in which the transcription matched one of the six stimulus syllables. The circle color indicates the stimulus for that syllable, while the color of the center "x" indicates what syllable was perceived. Three examples are highlighted. The blue-dashed box shows an error in which the stimulus was /du/, the perceived syllable was /bi/, and the kinematic data results in this production lying within the cluster of nonerror productions of /bi/. The red solid box shows a syllable where the stimulus was /za/ and transcribers perceived /da/, and the kinematics are on the boundary between the two categories. The black-dotted box highlights syllables in which the stimulus was /du/, but transcribers perceived /da/; the kinematics here are closer to the stimulus than to the perceived cluster.

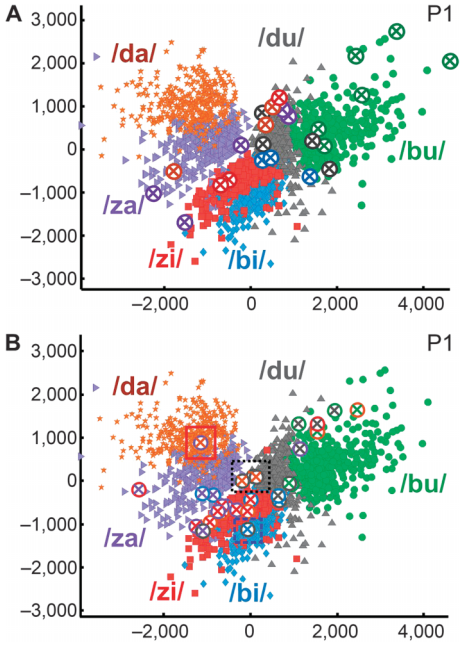
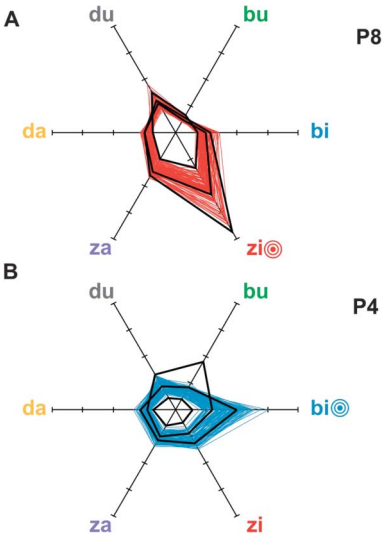


Figure 7. Similarity of distortion errors to canonical productions for the six syllables used in this study. Three contours indicating the similarity between productions transcribed as distortions and centroids of each of the syllable classes (each axis) are plotted with 100 randomly selected examples of nonerrors from the same stimulus class. The origin of each axis is 0 similarity (infinite distance from cluster centroid) and the end point is the highest similarity of any nonerror to the mean of all nonerrors in that class. (A) shows several distortions with the stimulus of /zi/ (target icon) from Participant 8 (P8) plotted in black on top of nonerror productions of the same class; these distortions all are most similar to /zi/. (B) shows three distortions of /bi/ from Participant 4 (P4) plotted on nonerror /bi/ productions; one production is similar to a canonical /bi/ (i.e., is very near the end point of the /bi/ axis), one is more similar to a canonical /bu/ than the nonerror /bi/ productions are, and one is dissimilar from all canonical classes (i.e., each distance from class centroids is large and thus near the origin of all axes).



canonical productions. Figure 8A shows a categorical error that had a stimulus of /du/ (target icon) and a perceived class of /bi/ (ear icon), plotted with randomly selected nonerror productions of both /du/ and /bi/. Note that this error primarily resembles canonical productions of the perceived class, /bi/. However, Figure 8B shows a categorical error that more closely resembles the stimulus class, /du/, than the perceptually determined class (/da/). In addition, Figure 8C shows a categorical error that had a stimulus class of /bu/ and a perceived class of /bi/; note that although this error is similar to both classes, the nonerror productions of these two classes are more distinct, suggesting this production may be a blend of the two articulation patterns.

Kinematic Traces of a Categorical Error Compared With Distortion and Nonerror Traces

While a detailed analysis of the specific kinematic abnormalities observed in error production is beyond the scope of the present study, we highlight a particularly interesting example that illustrates the complexity of characterizing DAF related effects. Figure 9 shows the articulograph (three sensors in one dimension) for an example categorical error plotted against kinematic traces for corresponding nonerrors and distortions. This error was classified as a syllable anticipation in which the intended stimulus sequence was /zabiduza/, and the participant instead produced /zabiza/. Thus, we compare this production to nonerror productions of the perceived sequence /biza/, the stimulus sequence /bidu/, and distortions of the /a/ in /biza/. Note that these distortions appear within the error bars (1 *SD*) provided by nonerror productions of the stimulus sequence until the final vowel. The trace of the sensor on the TD shows that the categorical error follows the perceived sequence /biza/ as expected (see Figure 9A).[3] The TT, however (see Figure 9B; red box), follows a movement trajectory appropriate for the stimulus sequence /bidu/ for some time before switching to

[3]Although there is a difference between the trajectories of the categorical error and the perceived sequence in the first 30 time points, this is consistent with initial coarticulation differences that are not indicative of an error; due to the organization of the stimuli, all nonerror sequences were /**u**biza/, whereas this error sequence was produced as /**a**biza/.

Figure 8. Similarity of three example categorical errors (in black) to all six canonical syllable classes, plotted with 100 randomly selected nonerror productions of their stimulus class (solid; target icon) and their perceived class (dashed; ear icon). Origin of axes is 0 similarity (infinite distance from cluster), and end point is the similarity of the nonerror that is most similar to the mean of all nonerrors in that class. (A) shows a categorical error from Participant 8 (P8) with a stimulus of /du/ and a perceived class of /bi/, plotted on top of nonerror productions of /du/ and /bi/; this error closely resembles canonical productions of the perceived class. (B) shows a categorical error produced by Participant 1 (P1) with a stimulus class of /du/ and a perceived class of /da/; this error closely resembles the stimulus class. (C) shows an error from Participant 7 (P7), which had the stimulus /bu/ and was perceived as /bi/; this error is more similar to both /bu/ and /du/ than nonerror productions.

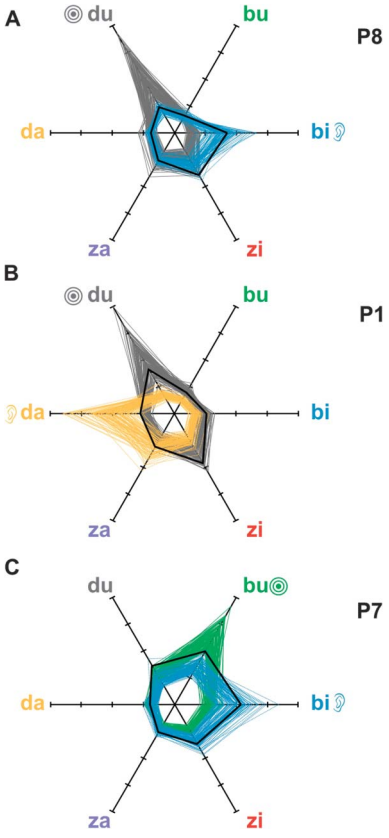
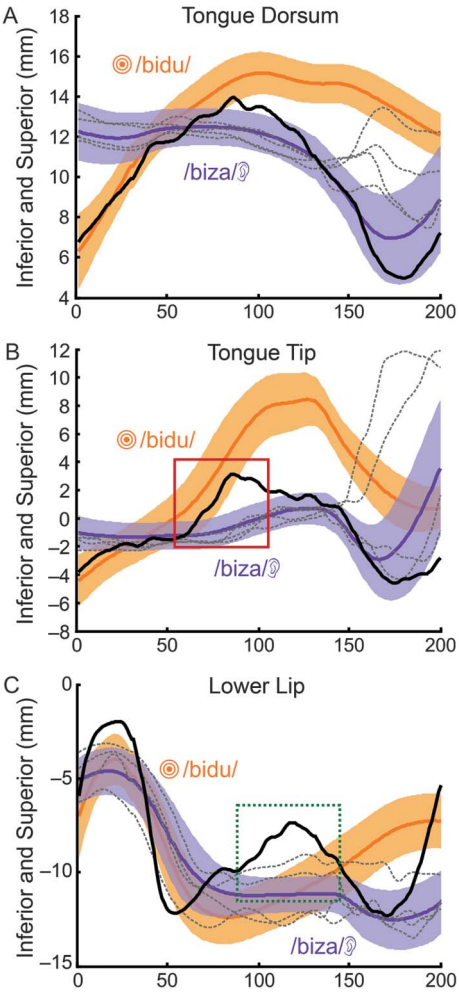


Figure 9. Kinematic traces for an illustrative categorical error plotted with traces for corresponding nonerrors and distortions. This categorical error, plotted in black in all three panels, was a syllable anticipation by Participant 6 (P6), in which the stimulus was /zabiduza/ and the participant instead produced /zabiza/. Three sensors are plotted showing the inferior and superior dimension. In orange is the mean ± 1 *SD* of all nonerror productions of the stimulus sequence /bidu/, whereas all nonerror productions of the perceived sequence, /biza/, are in purple. Distortions plotted in dashed gray lines were all intended productions of /biza/ with a perceived distortion on the final vowel. (A) shows the tongue dorsum position in the inferior and superior dimension, and (B) shows the tongue tip, in which the red box indicates where this sensor's movement initially matched the stimulus syllable and then veered into the perceived syllable. (C) shows the lower lip sensor, in which the green-dotted box indicates a time in which the lip movement does not appear to match either the stimulus or perceived nonerror sequences.



a trajectory that follows the trace for the perceived sequence /biza/. In the LL trace (see Figure 9C; green-dotted box), the trajectory of the categorical error does not clearly match either the perceived or the stimulus sequence.

Error Profiles

Errors were categorized within participant and then pooled to determine the distribution of the types of errors in the entire data set (see Table 4). Overall cross-validation classification accuracy of nonerrors, averaged across syllable classes within participant, varied from 91.5% to 99.7% by participant (M = 94.6%, SD = 2.5%). Table 4 shows the percentage of error productions that were sorted, on the basis of the responses of the six syllable-specific classifiers, into each of a set of error profiles. Of the 372 total distortion errors, 341 were transcribed as having one correct phoneme and one distorted phoneme and were submitted to the classifier. Of those, 47% were classified as matching the stimulus syllable only, 16% were classified as matching exactly one nonstimulus syllable, and 18% were classified

Table 4. Summary of error profiles across participants.

Distortions		Categorical	
Parameter	**Value**	**Parameter**	**Value**
Total number of errors	372	Total number of errors	143
One phoneme-matched stimulus[a]	341	Perceived as one of six canonical syllables[a]	81
Classified as[b]		Classified as[b]	
Stimulus only	47%	Stimulus only	5%
One other only	16%	Perceived only	43%
Stimulus plus one other	18%	Stimulus and perceived only	20%
Multiple	1%	Other combination or multiple	18%
None	18%	None	15%

Note. Shown are the total counts of errors of various types (across participants) and the percentage of errors that were classified into different error profiles using a quadratic discriminant analysis classifier.

[a]Productions that contained a combination of distortions and categorical errors (e.g., target /du/, transcribed /b@/) or in which the entire syllable was distorted (/@/) and categorical errors that were transcribed as a legal syllable but not one of the six in the entire stimuli set (e.g., /di/) were removed, as there were no canonical, nonerror productions to compare against. [b]Errors were classified via six classifiers per participant, each trained to recognize one syllable as in class and all other syllables as out of class. These were then used to create error profiles. For example, if a distortion error were classified as in class for only the stimulus syllable and out of class for the remaining syllables, it was denoted here in the left column as stimulus only.

as matching the stimulus and exactly one other syllable. The remaining syllables consisted of those classified as multiple syllables (1%) and those classified as not matching any canonical syllables (18%). Of the 143 total categorical errors, 81 were perceived as one of the six canonical syllables and thus could be directly compared with nonerrors; the remaining syllables were not contained in the stimuli (e.g., /ba/ or /di/) and thus could not be analyzed further. Of the syllables analyzed, 5% were classified as matching the stimulus syllable only, 43% were classified as matching the perceived syllable only, and 20% were classified as matching both the stimulus and perceived syllables only. The remaining syllables were either classified as a different combination of syllables and/or multiple syllables (18%) or classified as not matching any of the canonical syllables (15%).

Discussion

In this study, we developed and applied multivariate methods to examine the effects of DAF on articulatory kinematics of simple syllable sequences. We sought to first determine how, if at all, the kinematics of individual syllable productions judged by listeners to be correct (i.e., matching the stimulus) differed under DAF. Then, we compared syllable productions judged as errors (either distortions of the stimulus syllable or productions of a clean version of a nonstimulus syllable) to nonerror syllable productions from an articulatory perspective. The results, while preliminary, suggest the impact of altered timing of feedback on speech output can occur at either a subphonemic level (i.e., resulting in imprecise articulations) or at a categorical or phonological level (i.e., resulting in the selection and clean production of a nontarget syllable). We hypothesize that these qualitatively distinct error types are evoked by mechanisms that compare external auditory feedback with expectations either continuously (to determine if the articulation of a single syllable is correct) or discretely (to determine if the correct phonemes are being produced in the proper order). In the former case, artificially induced mismatches due to DAF may result in attempts to continuously adjust output of the current syllable, resulting in the observed distortions. In the latter case, errors (i.e., recognition of a planned sound at an unexpected time) may cause resetting or other changes in an abstract or phonological representation of the sequence in a speech-planning buffer, resulting in discrete sequencing errors. Because our study demonstrates variability in syllable productions that were not apparent to listeners, it also emphasizes the importance of studying articulatory kinematics to more precisely determine the nature of DAF-induced speech errors.

Nonerror Productions

The articulatory kinematics of nonerror syllable productions largely clustered by perceived syllable using a class-blind distance measure and nonmetric MDS (see Figure 4 and Supplemental Material S1). MDS plots were informative in demonstrating the global landscape of productions but are limited by the requirement to capture very high dimensional relationships in a two-dimensional space. Still, these plots illustrate a general (within-participant) consistency in articulating individual syllables within the repeated production of memory-guided CVCVCV sequences.

Five out of eight participants had statistically significant differences in the distribution of distances between individual productions and the cluster centroid for nonerror productions with and without DAF. For these participants, this suggests that articulations were still affected by DAF and were presumably less stable than those made with normal feedback, even when judged to be error free by a

listener. This builds on the results of Sasisekaran (2012), who found that lip aperture measured during the production of nonerror syllable sequences with DAF was more variable than those produced with typical or gated auditory feedback. The three participants who did not show statistically significant differences between kinematics for productions with and without DAF were those who made the lowest percentage of additional errors under delay (see Table 4). Although individual differences in susceptibility to DAF have been previously established (e.g., Burke, 1975; Chon et al., 2013), this result indicates that even subtle DAF-induced effects on the sensory-motor control of speech that are not apparent perceptually have differential effects on speakers; preliminary indications suggest that susceptibility to these subtle, graded alterations of articulatory output mirror susceptibility to more perceptually salient effects that clearly alter the output sound sequence. This suggests that the kinematic distance measure used here is sensitive to articulation changes under delay that were not evident to listeners, as all productions here were perceived as nonerrors.

These results were echoed by those using an adaptation of the well-established spatiotemporal index measure (Smith et al., 2000; here, ΔSTI is the difference in STI between delay and nondelay conditions); namely, those participants with the highest increase in error rates, as well as the highest difference in distribution of distances between individual productions and the cluster centroid also showed larger variability under delay rather than nondelay. There was one participant (P7) who showed a negative ΔSTI value but a statistically significant ($p < .05$) increase in variability under DAF using our new approach. This may indicate an increased sensitivity to articulatory variability using our method, but we should note that STI calculations were performed over CVCVCV sequences, while results shown in Table 4 were based on individual syllable productions.

Error Productions

Distortions

Errors judged to be distorted versions of the stimulus syllable tended to appear on the MDS plots near the stimulus (intended) syllable but often were further away from the cluster centroids than nonerrors (see Figure 6A). This reflects in many cases, as expected, the target syllables were likely correctly planned and released for execution but were altered during the articulation process due to mismatching continuous, external auditory feedback. When compared with nonerror productions via distances to class centroids (see Figure 7A) or via classification (see Table 4), distortions were often most similar to their stimulus syllable or to their stimulus plus one other canonical syllable, although examples in which distortions were more similar to a different canonical syllable (see Figure 7B), unlike all syllables (see Figure 7C), or alike more than one canonical syllable could also be found. Still, nearly half (47%) of the distortion errors, when submitted to classifiers trained on error-free productions of each stimulus, could only be considered to match the target stimulus syllable. Although speculative, this again suggests that a large fraction of these errors were altered by low-level interactions of feedback with the motor output system, resulting in graded, continuous changes to the articulatory output. Interpretation of other examples will require a detailed analysis of the articulatory movements in relation to the surrounding phonetic context and will likely show large variability across the small number of examples available.

Categorical Errors

Categorical errors often appeared on MDS plots (see Figure 6B) near the perceived syllable cluster, as expected if the participant had articulated a clean (but incorrect) syllable. Such productions represent examples in which the speech planning system appears to have selected and released the produced sounds in the improper order. Such serial ordering errors might resemble many normally occurring slips of the tongue, which can be, in part, explained by various models that include an abstract phonological representation of the forthcoming speech plan (Bohland et al., 2010; e.g., Dell, 1986; Hartley & Houghton, 1996; Vousden, Brown, & Harley, 2000). We hypothesize that these errors may arise due to mechanisms normally used for detection and correction of such discrete errors in speech output, which must enact changes to the speech planning buffer (i.e., a reordering of planned sounds) to interrupt and correct running speech. In the case of DAF, mistimed feedback may mimic the detection of such errors and artificially drive changes to the (phonological) speech planning buffer, directly resulting in serial ordering errors.

However, some syllables transcribed as categorical errors appeared on the border between the clusters of the perceived and stimulus cluster, and some appeared nearest the stimulus cluster despite being perceived as a clean version of a different syllable. To examine the kinematic relationships between categorical errors and canonical productions further, we generated spider plots (see Figure 8), which illustrate the degree of (dis)similarity between each production and the canonical (class centroid) version of each of the six syllables used in the experiment. These examples point to the possibility that individual components of the productions (i.e., movements of single articulators or coordinated articulatory gestures) might be in error, as has been reported for slips of the tongue elicited in a speeded repetition task with normal feedback aimed to elicit slips (Goldstein et al., 2007). In addition, examples with articulations with similarity to both the stimulus and perceived syllable could represent simultaneous coproductions of more than one planned segment (Pouplier & Hardcastle, 2005). In general, the developed visualization methods provide intuitive techniques to determine relationships between individual productions and to generate specific hypotheses that can be tested directly with the underlying EMA data.

We also summarized error profiles of categorical errors on the basis of classification analysis, across participants. We found that many syllables were classified as the

perceived syllable only (43%), again consistent with these being discrete phonemic sequencing errors. However, a relatively large fraction were classified both as the intended stimulus and perceived syllables but not as any of the other possible syllables (20%; see Table 4). These results suggest that, in fact, many errors represent blends of articulations of the stimulus and perceived sound. In addition to providing information about the interaction of DAF and articulatory planning, this result confirms that the articulation measures employed here add information about the nature of these errors that was not evident to listeners. One particularly compelling example of a syllable transcribed as a categorical error was shown in Figure 9. This example shows strong evidence that, at least in this example, some articulators follow trajectories consistent with the perceived syllable (see Figure 9A), while others combine trajectories of the perceived and stimulus syllable (see Figure 9B), and some articulators follow neither expected trajectory (see Figure 9C). This result points to the complexity of the effects of DAF-induced errors and of interactions between altered feedback and the speech controller. Although the effects of DAF have been studied for many decades, it is clear that they are still poorly understood.

Quantitative Measures for Kinematic Data Sets

Although the number of participants in this study was modest, the number of productions measured per participant was quite large (M = 2,939 syllables). Constraining the stimulus set to six syllables resulted in many more nonerror utterances of each type than are typically analyzed by using measures of articulatory variability such as STI (e.g., 10–15 utterances; Smith et al., 2000) The type of data collection, abbreviated transcription process and data processing approach, allows for comparison across many utterances and requires novel methods of visualization and analysis. Here, we have presented a variety of measures (high-dimensional representation of kinematic data as a feature vector; Manhattan distance from mean nonerror production as global measure of difference in articulation), associated visualizations (MDS and spider plots), and analyses (machine learning for classifying errors into distinct profiles). While STI is not typically used on errorful productions, the methods developed here can be directly employed to help characterize errors. Although the main aim of this study was to begin to determine how DAF affects syllable articulation, these methods could be applied to a variety of large kinematic data sets. Further development could also enable analysis across participants to characterize both group effects and individual differences.

Limitations

A variety of factors could have affected the results presented here in uncontrolled ways. Although participants all reported no history of speech, language, or hearing disorders, they did not undergo a hearing screening prior to participating and thus may have had different sensitivities to feedback alterations. To produce large numbers of similar syllables and elicited errors, the stimuli in this study consisted of repeating sequences of nonsense CV syllables. This was particularly necessary given the time constraints inherent in the use of EMA sensors, which can detach after some time, particularly from lingual placements. However, it is unknown how well these results would correlate to errors produced in conversational speech or how well they would generalize to words and syllables with different segmental structures. The sensor locations were chosen a priori to maximally differentiate between the six phonemes. These locations were similar but not identical to those subsequently published by Wang, Samal, Rong, and Green (2016), which suggested that UL, LL, TT, and a sensor somewhat posterior to our TD performed equivalently to a larger sensor set for classifying phrases, and thus may be preferable. In future studies, we plan to adopt these positions, in part to help standardize EMA protocols and enable comparisons across experiments.

The auditory-perceptual analysis presented here enabled transcribing the large numbers of syllables required for these analyses. However, this meant that productions labeled as correct were only transcribed by one listener, while productions that differed from the stimulus were transcribed by two or three listeners. This suggests that although we are confident that the productions labeled here as errors actually contained acoustically salient deviations from the targets, a small fraction of syllables marked as nonerrors may have also contained some perceptually discernable differences.[4] In addition, the kinematic measures presented here are intended to supplant the need to rely on auditory perception, as indeed different listeners may have different perceptual biases and thus may not identify the same set of errors. However, these kinematic measures still rely on manual identification of the syllable onset through listening to and viewing the recorded signal. Future development could use kinematic rather than perceptual markers to determine syllable boundaries. In a similar way, it is unknown to what degree the errors produced under DAF resemble those produced in typical speech or in studies eliciting slips of the tongue via rapid speech or tongue twister stimuli.

All analyses presented here were completed within-participant only, with the exception of the error profiles (see Table 4), which were calculated within participant but summed across participants. This is primarily because, although our sample size is consistent with some of the most directly relevant EMA studies (e.g., Goldstein et al., 2007; Tilsen & Goldstein, 2012), the effects of DAF are

[4]To help characterize the extent of this issue, the second transcriber performed a blind transcription of one entire run. Of the 668 syllables analyzed, there was exact agreement on 615 (92%). Where there was disagreement, most were marked as distortions by one transcriber and not the other (7% of total), rather than disagreements about error type (< 1% of total). Further, in nonerror syllables, the median jitter of the acoustically defined syllable boundary was less than 5% of the temporal window used for analysis.

in fact quite variable between participants. Therefore, the results presented here may not be representative of the population of typical speakers, and larger studies will ultimately be needed to assess individual variation of DAF effects on speech kinematics. In addition, the measures, visualization methods, and machine learning analyses have not been validated against other measures, with the exception of the presented STI comparison for nonerror data.

Future Directions

The complexity of this data set and of the methods presented here warrant further analysis. In particular, further study could reveal what specific articulatory changes shift a syllable from being perceived as a nonerror to a distortion, as we have presented evidence that DAF induces some changes in even clean productions in individuals most affected by feedback alterations. Further work could help collapse and compare data across participants, as most of the analyses presented here were within-participant only. The methods presented here were data driven, as variability in all sensors and dimensions were weighted equally; however, kinematic variation in some sensors and dimensions were almost certainly more critical than others for meeting phonetic goals. As the sensors and dimensions of import vary from phoneme to phoneme, the appropriate weighting of different sensors is not straightforward, and data-driven methods (e.g., on the basis of the Mahalanobis distance) may introduce circularity and/or biases in the analysis. However, future work should focus on weighting the different sensors and dimensions differentially on the basis of a priori knowledge of acoustic phonetics related to the specific target sounds used.

Also, these data and results provide evidence that can be used to develop and test models of speech planning and serial speech production. Although models have been developed that explicitly address the sequential representation and production of an utterance (e.g., Bohland et al., 2010; Tilsen, 2013), they differ in terms of the proposed planning units (i.e., abstract phonemes or articulatory gestures) and cannot currently explain how the mismatched timing of expected and received auditory feedback under DAF drives sequential production errors.

Conclusion

In this study, we examined the articulatory kinematics of syllable productions recorded from eight typical adult speakers with and without DAF. We show that analysis of articulatory kinematics can be informative in elucidating the precise nature of the speech errors made under DAF, which in many cases, differ from what is heard by listeners. We found that the kinematics of nonerror productions clustered by perceived syllable class using a data-driven distance measure operating on a global, high-dimensional representation of each production. Five participants had statistically significant differences in the variability of nonerror productions made with and without DAF, suggesting that kinematic measures are sensitive to articulatory changes under delay that are not evident to or not reported by listeners. Many distortion errors were observed to be most similar to the stimulus syllable, but others were similar to other syllables or alike no other syllables. Categorical errors were characterized in four groups: those that were most similar to the perceived syllable (most common), those that were most similar to the stimulus syllable, those that were a blend of the perceived and stimulus articulations, and those that were not similar to any of the six canonical syllables. The errors that are a blend of the perceived and stimulus articulations suggest that DAF may be interacting with the speech motor sequencing controller at a subphonemic level. This study is one of the first to characterize articulatory kinematic changes under DAF and provides initial results and a rich data set for testing future specific hypotheses and for constraining and testing computational models of speech planning and online control. In attempting to characterize these speech sequencing errors, we have developed and presented a series of novel quantitative methods for analyzing and visualizing large kinematic data sets.

Acknowledgments

The authors thank Mark Tiede for generously sharing his scripts to filter and rereference electromagnetic articulography data, Shanqing Cai and Joe Perkell for their consultation, Jessica Malloy for her work on previous related projects and early help with this study, and Timothy Streeter for his expertise in sound calibration. This research was supported in part by National Institutes of Health Grants: T90 DA032484 (fellowship for G. Cler, grant awarded to David Mountain from NIDA), F31 DC014872 (awarded to G. Cler from NIDCD), R03 DC012651 (awarded to C. Stepp from NIDCD), and P30 DC04663 (awarded to S. Colburn from NIDCD).

References

Black, J. W. (1951). The effect of delayed side-tone upon vocal rate and intensity. *Journal of Speech and Hearing Disorders, 16,* 56–60.

Boersma, P., & Weenink, D. (1996). Praat, a system for doing phonetics by computer (Version 3.4) [Computer software]. Retreived from http://www.praat.org

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience, 22,* 1504–1529. https://doi.org/10.1162/jocn.2009.21306

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Burke, B. D. (1975). Susceptibility to delayed auditory feedback and dependence on auditory or oral sensory feedback. *Journal of Communication Disorders, 8,* 75–96.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *The Journal of Neuroscience, 31,* 16483–16490. https://doi.org/10.1523/JNEUROSCI.3653-11.2011

Chon, H., Kraft, S. J., Zhang, J., Loucks, T., & Ambrose, N. G. (2013). Individual variability in delayed auditory feedback

effects on speech fluency and rate in normally fluent adults. *Journal of Speech, Language, and Hearing Research, 56,* 489–504. https://doi.org/10.1044/1092-4388(2012/11-0303)

Corey, D. M., & Cuddapah, V. A. (2008). Delayed auditory feedback effects during reading and conversation tasks: Gender differences in fluent adults. *Journal of Fluency Disorders, 33*(4), 291–305. https://doi.org/10.1016/j.jfludis.2008.12.001

Cornelisse, L. E., Gagne, J., & Seewald, R. C. (1991). Ear level recordings of the long-term average spectrum of speech. *Ear and Hearing, 12,* 47–54. https://doi.org/10.1097/00003446-199102000-00006

Cutler, A. (1981). The reliability of speech error data. *Linguistics, 19,* 561–582.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Fairbanks, G. (1955). Selective vocal effects of delayed auditory feedback. *Journal of Speech and Hearing Disorders, 20,* 333–346.

Fairbanks, G., & Guttman, N. (1958). Effects of delayed auditory feedback upon articulation. *Journal of Speech and Hearing Research, 1,* 12–22. https://doi.org/10.1044/jshr.0101.12

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30,* 139–162. https://doi.org/10.1006/jpho.2002.0176

Garnham, A., Shillcock, R. C., Brown, G. D. A., Mill, A. I. D., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19,* 805–818. https://doi.org/10.1515/ling.1981.19.7-8.805

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103,* 386–412. https://doi.org/10.1016/j.cognition.2006.05.010

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96,* 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Hartley, T., & Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. *Journal of Memory and Language, 35,* 1–31. https://doi.org/10.1006/jmla.1996.0001

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13,* 135–145. https://doi.org/10.1038/nrn3158

Hickok, G., Houde, J., & Rong, F. (2011, February). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69,* 407–422. https://doi.org/10.1016/j.neuron.2011.01.019

Hixon, T., Weismer, G., & Hoit, J. (2008). *Preclinical speech science*. San Diego, CA: Plural.

Hotopf, W. H. N. (1983). Lexical slips of the pen and tongue: What they tell us about language production. In B. Butterworth (Ed.), *Language production, vol. 2: Development, writing, and other language processes* (pp. 147–199). London: Academic Press.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5,* 1–14. https://doi.org/10.3389/fnhum.2011.00082

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115–129. https://doi.org/10.1007/BF02289694

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research, 14,* 677–709. https://doi.org/10.1044/jshr.1404.677

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences, 22,* 1–38.

Malloy, J. R., Nistal, D., & Bohland, J. W. (2014). *A study of speech sequencing errors due to delayed auditory feedback*. Paper presented at the Motor Speech Conference. Sarasota, FL.

Nooteboom, S. G. (1973). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech Errors as Linguistic Evidence* (pp. 144–156). The Hague, the Netherlands: Mouton.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 87–95), New York, NY: Academic Press.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59*(2), 441–449.

Perkell, J. S., Denny, M., Lane, H., Guenther, F., Matthies, M. L., Tiede, M., . . . Burton, E. (2007). Effects of masking noise on vowel and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users. *The Journal of the Acoustical Society of America, 121,* 505–518. https://doi.org/10.1121/1.2384848

Perkell, J. S., Numa, W., Vick, J., Lane, H., Balkany, T., & Gould, J. (2001). Language-specific, hearing-related changes in vowel spaces: A preliminary study of English- and Spanish-speaking cochlear implant users. *Ear and Hearing, 22,* 461–470.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77,* 97–132.

Pouplier, M. (2007). Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech, 50,* 311–341. https://doi.org/10.1177/00238309070500030201

Pouplier, M., & Goldstein, L. (2005). Asymmetries in the perception of synthesized speech. *Journal of Phonetics, 33,* 47–75. https://doi.org/doi:10.1016/j.wocn.2004.04.001

Pouplier, M., & Hardcastle, W. (2005). A re-evaluation of the nature of speech errors in normal and disordered speakers. *Phonetica, 62,* 227–243. https://doi.org/10.1159/000090100

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America, 119,* 2288–2297. https://doi.org/10.1121/1.2173514

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382. https://doi.org/10.1207/s15326969eco0104_2

Sasisekaran, J. (2012). Effects of delayed auditory feedback on speech kinematics in fluent speakers. *Perceptual and Motor Skills, 115,* 845–864. https://doi.org/10.2466/15.22.PMS.115.6.845-864

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 109–136). New York, NY: Springer-Verlag.

Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18,* 41–55. https://doi.org/10.1016/S0022-5371(79)90554-1

Smith, A., Johnson, M., McGillem, C., & Goffman, L. (2000). On the assessment of stability and patterning of speech movements. *Journal of Speech, Language, and Hearing Research, 43,* 277–286. https://doi.org/10.1044/jslhr.4301.277

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America, 111,* 2237–2241. https://doi.org/10.1121/1.1466868

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America, 84,* 917–928. https://doi.org/10.1121/1.396660

Tiede, M., Bundgaard-Nielsen, R., Kroos, C., Gibert, G., Attina, V., Kasisopa, B., . . . Best, C. (2010, 15–19 November). Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously. *Proceedings of Meetings on Acoustics*. Paper presented at 160th Meeting Acoustical Society of America, Cancun, Mexico. https://doi.org/10.1121/1.3508805

Tilsen, S. (2013). A dynamical model of hierarchical selection and coordination in speech planning. *PloS One, 8*(4), e62800. https://doi.org/10.1371/journal.pone.0062800

Tilsen, S., & Goldstein, L. (2012). Articulatory gestures are individually selected in production. *Journal of Phonetics, 40,* 764–779. https://doi.org/10.1016/j.wocn.2012.08.005

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage, 39,* 1429–1443. https://doi.org/10.1016/j.neuroimage.2007.09.054

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America, 122,* 2306–2319. https://doi.org/10.1121/1.2773966

Vousden, J. I., Brown, G. D., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology, 41,* 101–175. https://doi.org/10.1006/cogp.2000.0739

Wang, J., Samal, A., Rong, P., & Green, J. R. (2016). An optimal set of flesh points on tongue and lips for speech-movement classification. *Journal of Speech, Language, and Hearing Research, 59,* 15–26. https://doi.org/10.1044/2015_JSLHR-S-14-0112

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America, 116,* 1168–1178.

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin, 60,* 213–232. https://doi.org/10.1037/h0044155