Delayed auditory feedback elicits specific patterns of serial order errors in a paced syllable sequence production task

Jessica R. Malloy[a], Dominic Nistal[b], Matthias Heyne[c], Monique C. Tardif[cd], and Jason W. Bohland[cd*]

a. Program in Neuroscience, Boston University, Boston, MA 02215, USA
b. University of Washington, Department of Neurological Surgery, Seattle, WA 98104, USA
c. Department of Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA 15260, USA
d. Center for the Neural Basis of Cognition, Pittsburgh, PA 15213, USA

* Corresponding author

j.bohland@pitt.edu

Keywords: speech, speech production, speech planning, feedback, delayed auditory feedback, sensorimotor integration, modeling

Malloy, J. R., Nistal, D., Heyne, M., Tardif, M. C., & Bohland, J. W. (2022). Delayed Auditory Feedback Elicits Specific Patterns of Serial Order Errors in a Paced Syllable Sequence Production Task. *Journal of Speech, Language, and Hearing Research : JSLHR*, *65*(5), 1800–1821. https://doi.org/10.1044/2022_JSLHR-21-00427.

Abstract:

Purpose: Delayed auditory feedback (DAF) interferes with speech output. DAF causes distorted and disfluent productions and errors in the serial order of produced sounds. While DAF has been studied extensively, the specific patterns of elicited speech errors are somewhat obscured by relatively small speech samples, differences across studies, and uncontrolled variables. The goal of this study was to characterize the types of serial order errors that increase under DAF in a systematic syllable sequence production task, which used a closed set of sounds and controlled for speech rate.

Method: Sixteen adult speakers repeatedly produced CVCVCV sequences, paced to a “visual metronome,” while hearing self-generated feedback with delays of 0 to 250 ms. Listeners transcribed recordings, and speech errors were classified based on the literature surrounding naturally occurring slips of the tongue. A series of mixed-effects models were used to assess the effects of delay for different error types, error arrival time, and speaking rate.

Results: Delayed auditory feedback had a significant effect on overall error rate for delays of 100 ms and greater. Statistical models revealed significant effects (relative to zero delay) for vowel and syllable repetitions, vowel exchanges, vowel omissions, onset disfluencies, and distortions. Serial order errors were especially dominated by vowel and syllable repetitions. Errors occurred earlier on average within a trial for longer feedback delays. While longer delays caused slower speech, this effect was mediated by the run number (time in the experiment) and small compared to previous studies.

Conclusions: Delayed auditory feedback drives a specific pattern of serial order errors. The dominant pattern of vowel and syllable repetition errors suggests possible mechanisms whereby DAF drives changes to the activity in speech planning representations, yielding errors. These mechanisms are outlined with reference to the GODIVA model of speech planning and production.

1 Introduction

It has long been known that delayed auditory feedback (DAF) can have profound impacts on speech motor control (Lee, 1950, 1951). The auditory playback of a speaker’s own vocal output with a temporal lag leads to reductions in speech rate, increases in vocal intensity and fundamental frequency (Fairbanks, 1955), and an array of errors in fluency (Yates, 1963). Despite an extensive literature using this technique in both healthy speakers and individuals with fluency disorders (e.g., De Andrade &

Juste, 2011; Kalinowski, Armson, Stuart, & Gracco, 1993; Kalinowski, Stuart, Sark, & Armson, 1996; Soderberg, 1969), the mechanisms responsible for the observed errors remain relatively poorly understood. This may be, in part, due to the bluntness of the manipulation, which can produce particularly strong, global, and seemingly chaotic effects. Indeed, Cai et al. (2011) noted that such "gross, nonspecific alterations are of limited value in understanding speech under ordinary circumstances."

Driven in part by these concerns, research in sensory-motor control of speech has recently focused more prominently on the effects of *frequency-altered feedback* (FAF), in which fundamental frequency or specific spectral components (e.g., formant frequencies) of an utterance are modified and presented back to the participant in near real-time (e.g., Burnett, Freedland, Larson, & Hain, 1998; Purcell & Munhall, 2006). In such manipulations, speakers typically make small, compensatory (but highly variable) changes in their speech, usually explained as an effort to steer auditory feedback in order to better match some learned auditory expectations for the sound(s) being produced. These modest changes can be described as *analog* or *continuous* modifications of speech output, typically not crossing sound category boundaries or involving transpositions or substitutions in the sequence of sounds that are produced. On the other hand, many of the errors that have been noted previously in DAF experiments appear to involve *discrete*, *categorical* modifications in the sequential output of speech. Thus, although both FAF and DAF effects are driven by neural processes involving the incoming feedback signal, they appear to elicit some fundamentally different responses. One intriguing possibility is that these manipulations may preferentially tap into different hierarchical levels of the sensory-motor circuitry that is used to guide and monitor speech. In this study we aimed to better understand the effects of DAF on speech output by developing a systematic, modernized, interpretable protocol and controlling for factors that are likely to mask the effects that DAF has on the control of serial speech. Our central hypothesis was that temporally induced mismatches between auditory expectations and incoming feedback would elicit large "error signals" that can drive predictable changes to the forthcoming speech plan, which are observable as speech output errors. Although DAF certainly elicits changes in the detailed acoustics of sound productions, our approach focused primarily on patterns of discrete speech errors, reflecting changes in the *sequential* output of speech, rather than on acoustic (i.e., spectral) analysis.

Among the most consistent results in the literature surrounding DAF is that subjects *reduce their speaking rates* under delayed feedback. The amount of speech rate reduction (or duration increase) depends on the delay interval. Early studies suggested that speech rate was maximally reduced when the delay was ~180-200 ms (ATKINSON, 1953; Black, 1951; Fairbanks, 1955). Slowing speech output may provide speakers a mechanism that allows them to, at least partially, account for the misalignment of expected and observed auditory feedback. That is, while reducing speech rate under DAF will not eliminate temporal mismatches, prolonging vowel sounds may offer the auditory system a *glimpse* of the expected sound within a temporally constrained processing window. For models of speech production that incorporate an auditory feedback processing circuit (e.g., Guenther, 2016; Guenther, Hampson, & Johnson, 1998; Hickok, Houde, & Rong, 2011; Houde & Nagarajan, 2011), this could have the effect of reducing sensory mismatch or error[1]. If the minimization of sensory error is part of an overall control strategy, then reducing rate under DAF may be seen as a compensatory mechanism that targets acoustic-phonetic control variables. Rate reductions, however, could also occur for other reasons; for example, interference from the mismatch between feedback and expectations might engage additional cognitive processes, resulting in slowing due to a “bottleneck” of neural resources.

Acoustic-phonetic level feedback control models cannot clearly account for the other substantial effects that are observed in studies using DAF. Speakers, in many cases, are relatively unable to maintain typical fluency in the face of delayed auditory signals. This is in contrast to, for example, pitch-shifted feedback, in which subjects either compensate for or follow the direction of the pitch-shift (Behroozmand, Korzyukov, Sattler, & Larson, 2012; Burnett et al., 1998) by making small, online adjustments to the controller without disruptions in fluency, even in running speech (Patel, Niziolek, Reilly, & Guenther, 2011). DAF, on the other hand, causes a range of disfluencies, which include serial ordering errors that are not dissimilar from those that occur naturally at a vastly lower frequency. Such slips are thought to be subject to *error monitoring* processes that may occur at multiple levels of the linguistic hierarchy and may involve the operation of both internal (via forward models) and external (via sensory feedback) monitors (Levelt et al., 1999; Postma, 2000).

What *are* the types of *errors* that speakers make when subjected to DAF? Considerable historic emphasis has been placed on the so-called “artificial stutter” (Lee, 1950, 1951), originally described as taking the form of an undesired repetition of syllables or fricatives (Lee, 1950). In early work on this topic, Fairbanks and Guttman (1958) classified “substitution,” “omission,” and “addition” errors, and noted that “the most distinctive characteristic of peak disturbance is high incidence of additions,” of which they noted ~70% were repetitions of sounds. Such errors were most prevalent at a delay interval of 200 ms, increasing ~20-fold over the normal auditory feedback (NAF) condition. In a more recent investigation, Chon et al (2013) studied a relatively large number of healthy speakers (*N*=62), with a focus on individual variability. They classified speech errors that included omissions, substitutions, and

[1] If a participant spoke under 200 ms delay, for example, producing vowels with greater than 200 ms duration would allow feedback the initial portion of the vowel production to arrive at the ear before the speaker completes production of that vowel. Since monophthong vowels have relatively steady-state acoustics, this strategy would typically result in a close match between auditory feedback and speaker expectations for portions of an utterance.

additions in spontaneous conversational speech under NAF and 250ms delayed feedback. These discrete errors occurred much more commonly (~1-1.5 errors per 100 syllables) under DAF than NAF. Likewise, “stuttering-like disfluencies” (part-word and single-syllable word repetitions; Ambrose & Yairi, 1999) were much more prominent, on average, under DAF. Unfortunately, no more specific breakdown of these errors by type was available from this study. A recent exploratory study investigating articulatory kinematics (Cler, Lee, Mittelman, Stepp, & Bohland, 2017) used a paced syllable sequencing paradigm, in which participants were encouraged not to reduce rate, and found a preponderance of errors deemed by listeners to be discrete sound errors including sound repetitions. Through examination of articulatory data, it appears that these represent both discrete errors and errors likely to involve co-productions or blends of individual syllable productions.

Altered feedback studies consistently provide evidence for the use of auditory feedback in online speech motor control. Discrete, categorical errors in the serial output of speech (i.e., errors involving whole sound repetitions, omissions, or substitutions) argue for a “higher-level” interaction involving a speech planning buffer, where items are incorrectly (and often repeatedly) selected for output by the production system. The GODIVA model of speech sequencing (Bohland, Bullock, & Guenther, 2010) offers one neurocomputational explanation for this planning buffer and selection process but cannot currently account for the types of delay-induced error patterns described above. In order to extend this model to simulate these higher-level perception-production interactions and account for effects such as those observed under DAF, it is important to develop a systematic, quantitative dataset describing speech under different temporal feedback delays.

In this study, we conducted a carefully controlled investigation of the error patterns that emerge in a nonword syllable sequence production task under DAF. We tested the hypothesis that increasing feedback delays (and in turn increasing discrepancies between auditory expectations and feedback) would yield a systematic, non-random pattern of discrete speech errors. Such findings would provide important clues to help constrain an expanded model of speech sequencing that incorporates error monitoring and auditory feedback. We specifically controlled speech rate to better isolate the effects that are directly related to mismatched timing within the control circuit independent of rate-related compensations, which serve to reduce these discrepancies. Specifically, if reduced speech rate is a compensatory strategy aimed at reducing “low-level” mismatch between auditory expectations and external auditory feedback, we argue that this behavior may also mask the impact of DAF on the normal function of speech control loops at multiple levels. We hypothesized that this paced speech approach could be expected to accentuate and amplify sequencing errors, providing a critical mass of such errors to enable their precise quantification. More specifically, we expected that as the delay magnitude approached the time between syllable production onsets, auditory “error signals” would be maximized. If, as we hypothesized, these error signals can drive direct changes to the representation of the forthcoming speech plan, then we would expect to observe large numbers of discrete sound errors as feedback delays increased. If, as we propose, these error-driven modifications to speech output representations are non-random (i.e., not simply driven by an overall increase in "noise"), then we would expect some error types to be impacted more than others. To address these general hypotheses, we collected a large sample of utterances from each participant, with each non-lexical stimulus composed from the same small, closed set of sounds produced under six different feedback latencies. Manual transcriptions combined with automated analysis of speech errors address our general hypotheses and provide a new perspective on sound sequencing errors under DAF.

The overall goal of this study was to test the hypothesis that delayed auditory feedback drives non-random speech serial order errors using a highly controlled setting. Results were expected to inform and eventually improve theoretical models of speech motor control. The specific research questions we sought to address were as follows:

- Based on a classification using methods drawn from the literature on naturally occurring speech errors, which specific error types are observed more frequently under increasing auditory feedback delays?
- Do such serial order errors arise earlier within a trial for longer delays than for shorter delays, which would suggest an accumulating effect of error on changes in speech output?
- Can an external timing signal help participants resist the tendency to reduce speech rate under DAF, essentially making a tradeoff of increased speed for decreased accuracy?

2 Method

2.1 Participants

16 college-age participants (9 women) took part in the study. All participants were right-handed, first-language American English speakers and self-reported no history of speech, language, or hearing disorders or other neurological conditions. Informed consent was obtained from all participants in accordance with the Boston University Institutional Review Board. One female participant's data were removed from the analysis due to a consistent failure to produce the intended stimulus. One male participant's data were additionally excluded due to problems following the experimenter's instructions. Thus, data from 14 participants (8 women) were included in the final analysis.

2.2 Experimental setup

Figure 1 illustrates the basic experimental setup used in this study. Participants were seated comfortably inside a soundproof booth in front of an LCD monitor, which displayed experimental stimuli and other visual cues. Stimulus delivery was controlled using the Psychophysics Toolbox (PTB-3) for MATLAB (Mathworks, Inc; Natick, MA). Participants were fitted with closed circumaural headphones (Sennheiser HD280 Pro; Wedemark, Germany), which provided up to 32 dB ambient noise attenuation, and a head-worn microphone (Shure WH30XLR Cardioid Condenser Microphone; Niles, IL) positioned approximately 4 cm from the corner of the mouth.

Speech signals were transmitted from the microphone to an external sound device (M-Audio Fast Track Ultra; Cumberland, RI) connected to a laptop computer via USB-2.0. Auditory feedback delays were specified using PsychPortAudio, a software sound interface available in the Psychophysics Toolbox (PTB-3), which utilizes Audio Stream Input / Output (ASIO) drivers to obtain high temporal precision and low latency sound playback using specialized audio devices. A distinct mode of processing ("ASIO Direct Monitoring") was used for "zero" delay trials in order to provide feedback with the lowest latency achievable by the hardware (~6-7 ms measured using the method described by Kim, Wang, & Max, 2020). The M-Audio sound device transmitted the (delayed) speech signal to the participant's headphones via a mixer/amplifier (Behringer Xenyx 802; Branchville, NJ). The overall system was set to achieve (for all delays including zero) an approximate +5dB gain from the vocal signal measured at the microphone to the signal output measured at the headphones. The participants' speech (microphone signal) was recorded digitally with sampling rate of 44.1 kHz using Audacity software (http://audacityteam.org) on a secondary computer for offline analysis.

2.3 Experimental task

The experimental task was to repeatedly produce memory-guided speech sequences under various levels of delayed auditory feedback. Speech stimuli consisted of 3-syllable CVCVCV sequences comprised of the stop consonants /b/, /d/ and /g/, and point vowels /a/, /i/, and /u/. Each of the six phonemes appeared in every stimulus (i.e., no phonemes were repeated), and all possible combinations occurred with equal probability for each participant over the course of the experiment.

On each trial in each of 5 runs, one sequence – chosen pseudorandomly from the full set, which was identical for each run – was presented orthographically (e.g., "boo dah gee") on the monitor for 3 seconds. The subject was asked to memorize and prepare to repeatedly produce the memorized sequence when cued. The cue to speak was a visual change, in which the stimulus was removed and replaced by a "visual metronome" that encouraged subjects to produce steady, consistent, rhythmic speech, with target rate of 5 Hz (200 ms between syllable onsets). The visual metronome took the form of three circles positioned horizontally on the screen (see Figure 1, top), which sequentially changed colors from white to green at the target rate, and which remained on screen for the duration of a 5 second production period.

Subjects were instructed to overtly produce the most recently presented sequence *repeatedly* throughout the production period while receiving auditory feedback through headphones, and to pace their utterances approximately to the metronome (i.e., to time the onset of each syllable to a color change). Subjects were directed to attempt to maintain the metronome rate as closely as possible, even to the detriment of "correct" speech output if necessary. These instructions were intended to combat the natural tendency for speakers to reduce their speech rate under DAF.

Each experimental run consisted of 60 trials, each involving presentation and production of one sequence. Subjects were asked to complete 5 runs, each of which was ~11 minutes in duration. Of the 14 participants who were included in the final analysis, 12 completed 5 runs and 2 completed 4 runs. During each run, the delay for each trial was chosen pseudorandomly between 0[2] and 250 ms in 50 ms intervals (i.e., {0, 50, 100, 150, 200, 250 ms}) to reduce subject adaptation to a specific delay and to test the effects of different delays. Zero delay trials were more than twice as frequent as any specific non-zero delay, occurring 20 times per run, while each non-zero delay occurred 8 times per run. This manipulation was designed to lower subject stress and fatigue, which can be high when forced to speak under conditions that frequently cause disfluency. 5 of the 20 zero delay trials were placed at the start of each run, in order to allow subjects to acclimate to the task of speaking with auditory feedback via headphones.

2.4 Speech analysis

Across participants, we analyzed data from 4080 trials. Audio files were parsed into individual trials, each of which was manually transcribed by one of five individuals using a custom machine-readable format. The process required the transcriber to chart perceived phonemes across the 5 s utterance; because the 6 phonemes used were the same for all utterances, transcribers were asked to indicate instances of those phonemes and to mark all other phonemes or severely distorted versions of a phoneme using a special character. Transcribers also noted the total duration of the utterance (i.e., from the onset of the first syllable produced through the offset of the last syllable *transcribed*), using Audacity software.

Custom MATLAB software was written to process the transcription files. Each trial was automatically annotated as either a correct production (i.e., the subject produced the correct target

[2] As noted above, the "zero" delay trials had the smallest delay possible using this hardware, of approximately 6-7 ms. When we refer to zero delay trials, it should be understood that these are actually minimum delay trials.

sequence in all instances without any clear errors) or an error production, in which either a sound sequencing error or severe sound distortion occurred. To further annotate error productions, the software classified the *first* speech sound error made in the produced sequence; only the first error was analyzed because it was impossible to determine whether subsequent errors occurred independently from, or as a direct result of, the first error, making interpretation exceedingly difficult. Furthermore, even classifying subsequent syllables produced as correct or in error is made difficult because participants may "reset" their position within the sequence as a result of detecting an error. Errors were classified as one of 6 major types: *repetitions*, *anticipations*, *exchanges*, *omissions*, *onset disfluencies*, and *distortions*[3]. Repetitions, anticipations, and exchange errors were further divided into three variants – consonant, vowel, and syllable – that describe the element in error, resulting in 13 distinct error types. This choice of error classification scheme was based on several considerations, including: (i) a well-developed literature concerning sound movement errors in typical speech, providing the basic categories used; (ii) the ability to relate these errors to our GODIVA modeling framework (Bohland et al., 2010); and (iii) the ability to provide a relatively unambiguous classification of errors in our highly controlled repetitive syllable production task. The 13 error types studied (see Table 1 for a more detailed description and example of each) do not represent all changes to speech output that could be or were elicited as a function of feedback delay. Disfluencies could also include blocks and prolongations, which were not explicitly included in our error classification scheme. Additionally, modifications in pitch, intensity, or other small acoustic changes were not specifically addressed.

The software also output which syllable in the sequence was the first in error (i.e., which syllable was used for error classification), and the average inter-syllable duration for each trial, defined as the duration noted by the transcriber divided by the number of syllables in the trial. Trials containing errors that occurred within the first three syllables (i.e., within the first utterance of the sequence) but which formed a legal sequence in this experiment (i.e., a CVCVCV sequence using the 6 eligible phonemes with no repeats) were considered failures to recall the presented sequence and not a speech sequencing error. A total of 123 trials (~3% of the total) met these criteria and were excluded from further analysis.

2.4.1 Statistical models of error occurrence

Error data were modeled using generalized linear mixed effects models implemented in the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R version 4.0.3. The *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) was used for testing statistical significance, and the

[3] Onset disfluency is a term we adopt in place of "artificial stutter" (Lee, 1951) to describe errors made that involved repeated productions of the onset consonant without a clear intervening vowel. Distortions were limited to relatively severe changes in acoustics that were not obviously an example of the target sound.

emmeans package (Lenth, Singmann, Love, Buerkner, & Herve, 2018) was used for *post hoc* comparisons.

To analyze the effects of delay on speech errors, we used mixed effects logistic regression to model binary outcome variables (presence or absence of an error of a given type). The *glmer* function was used, with family set to *binomial*; the *BOBYQA* algorithm (Powell, 2009) was used for optimization. To select an appropriate model, we tested models with different variables included / excluded as fixed and random effects for predicting the presence or absence of *any* error (trials were coded as 1 if any error occurred, and 0 if no error occurred). The model with lowest Bayesian Information Criterion (BIC) was selected. Predictor variables tested included *delay* (0, 50, 100, 150, 200, 250 ms; as a categorical variable), *duration* (average inter-syllable onset duration for each trial), *run* (indexed sequentially across the session), *sequence* (CVCVCV item for each trial; as a categorical variable), and *subject* (as a categorical variable). Dummy coding was used for all categorical variables (e.g., different levels of the *delay* variable were compared to the reference level of zero). Models that included a random slope for the factor *delay* did not, in general, converge, and were eliminated. Random intercepts were included for *subject* and *sequence*; eliminating either increased the BIC. The selected model included *delay* as a fixed effect, with random intercepts for *subject* and *sequence*:

$$error \sim delay \; + \; (1 \mid subject) + (1 \mid sequence).$$

Visual and statistical model diagnostics were performed using the DHARMa package (http://florianhartig.github.io/DHARMa/). The selected model structure was then applied repeatedly to test for effects of delay on *individual* error types (see Table 1). Post-hoc multiple comparisons (pairwise for the six levels of delay 0 – 250 ms) used the Tukey's posthoc method for comparing a family of six estimates.

2.4.2 Statistical model of error arrival time

A mixed effects log-linear regression model was implemented to test for effects of delay on error arrival time, quantified as the number of syllables produced when the first error occurred, including the erroneous syllable. Trials that were determined to be error free were excluded from this analysis. This model was constructed with the same structure as those described above and was estimated using the *glmer* function. However, the distribution family was specified as Poisson, with a log link function. The form of the model was specified as:

$$arrival\ time \sim delay \; + \; (1 \mid subject) + (1 \mid sequence).$$

2.4.3 Statistical model of speaking rate

An additional linear mixed effects model was fit using the *lmer* function in *R* to determine the effects of delay, run, and trial number (i.e., time within a run) on syllable duration (a proxy for speaking rate). An additional fixed effect, *err*, was included to indicate trials that contained (any type of) errors. The intention of this model was to determine if speaking rate showed systematic variance explainable by these experimental factors, particularly delay. The dependent variable, *duration*, was the average inter-syllable onset duration in each trial. Because the first 5 trials in each run were always zero latency, these trials were excluded in the rate analysis. The model included random intercepts for subject and sequence (as above). Based on lowest BIC, we chose a model that excluded trial number and included an interaction between run and delay. In this model, run and delay are treated as categorical predictors (factors). Models using these terms as continuous predictors did not yield qualitatively different results. The form of the model was specified as:

$$duration \sim 1 + \ run * delay + err + (1 \mid subject) + (1 \mid seqs)$$

2.4.4 Temporal measures of repetition errors

A detailed exploratory timing analysis was performed for all trials that included syllable or vowel repetition errors (see Table 1). The Vocal Toolkit plugin (Corretge, 2020) for Praat v6.1 (Boersma & Weenink, 2001) was used to automatically generate TextGrid objects that segmented the recorded microphone signal into syllables and silences and marked syllable nuclei (de Jong & Wempe, 2009). Two of the authors, as necessary, manually adjusted these annotations to optimally mark syllable boundaries and vowel nuclei in the two relevant syllables (the first and second instances of the syllables involved in the repetition error). Syllable onsets were defined as the first release burst in the relevant acoustic signal (note all syllables began with voiced stop consonants), and syllable offsets were defined by the decrease in energy and disappearance of formant frequencies in the spectrogram.

From these annotations, three temporal measures were calculated (see Figure 2). First, *overlap* was defined as the proportion of the production of the second syllable (containing the repetition) that occurred while feedback of the first syllable was present. Second, the *onset interval* was defined as the time between the onset of the *feedback* from the first syllable and the onset of the *production* of the second syllable (positive if the onset of feedback from the first syllable *preceded* the onset of the production of the second syllable). The *peak interval* was defined as the time between the syllable nucleus of the feedback from the first syllable to the production of the syllable nucleus of the second syllable (positive if the time of the peak of the feedback from the first syllable preceded the produced peak in the second syllable). The distributions of these empirical measures were examined, and each measure was modeled using linear mixed-effects models with random intercept for *subject* and fixed effects of *delay* and *error type* (syllable or vowel repetition) and the interaction of these two variables.

$$temporal\ measure \sim delay * error\ type + (1 \mid subject)$$

The measure *overlap* is restricted to [0,1], so these values were first transformed using the *logit* function prior to model fitting. Since many overlap values were equal to 1.0, and would thus be transformed to ∞, values were remapped (via simple threshold) to [0.025, 0.975] prior to application of the logit transform.

3 Results

Audio recordings from 14 subjects, each of whom completed 300 trials (except S1 and S6, who completed 240 trials), were manually transcribed by a first trained listener. These transcriptions were compared against the target sequence and the first error, if any were made, was categorized (see Table 1). For any trials in which any error was found to occur, a second individual also provided a transcription. For any trials in which the error type was not agreed upon by the first two reviewers (as categorized by our custom software), a third individual provided a final transcription. Only data for which there was agreement on the first error type between at least two transcribers were considered further. A total of 244 trials were discarded due to disagreement among transcriptions (~6.0% of total trials).

3.1 Errors as a function of delay

Figure 3 illustrates the frequency of occurrence of different error types made by each participant as a function of delay. Because only a total of 5 errors were classified as consonant omissions, these were not included in any analyses. All participants made speech errors, though the incidence of errors differed greatly across delay intervals and participants. For example, Subject S8 made errors in more than 60% of trials with 250 ms delay, while S7 made errors in fewer than 10% of trials at the same delay. A further breakdown of each participant's errors into discrete substitution and non-substitution errors is provided in Supplementary Figures 1 and 2. We first tested the hypothesis that the probability of *any error* (e.g., the height of the stacked bars in Figure 3) is significantly modulated by the delay interval. As described in the *Methods* above, we tested several generalized linear mixed effects logistic regression models and selected the model with lowest BIC for further analysis. This model included a fixed effect for *delay* and random intercepts for *subject* and *sequence*. The model demonstrated significant effects of delay on the probability of a speech error for delays of 100 ms or greater ($p < 0.001$; see Table 2). Pairwise contrasts (p-values adjusted using Tukey's method) showed that all delays of 100 ms or more caused a statistically significant increase in error rate when compared to zero delay trials ($p < 0.001$), with odds ratios (OR; odds of error at a given delay compared to zero delay) increasing monotonically with delay. The odds of an error occurring with 250 ms delay were 7.18 (95% CI = 5.53 – 9.31) times higher at 250 ms delay than at zero delay. Pairwise contrasts showed that successive increases in delay caused a significant increase in error probability for 50 to 100 ms

(OR=1.67; *p=0.0485*) and from 100 to 150 ms (OR=1.84; *p = 0.0012*). Pairwise contrasts of 150 vs. 200 ms and 200 vs. 250 ms were not significant; however, error probability was higher for 200 ms delay than for 100 ms (or lower), and error probability was higher for 250 ms delay than for 150 ms delay (or lower).

3.2 Error patterns by type

Speech error results are further summarized in Figure 4, which shows the mean probability (across participants) of each error type occurring as the first error in a trial as a function of delay, grouped into repetition, exchange, anticipation, and other error types. Here it is apparent that *specific* error types (and not others) are elicited with increasing likelihood as the delay interval increases. To test the hypothesis that the probability of occurrence of each individual error type was modulated by the delay interval, we computed generalized linear mixed effects models with fixed effect *delay* and random intercepts for *subject* and *sequence* (the same model structure as above). Individual model summaries (including pairwise comparisons for different delays) are provided in the Supplemental Materials. To account for multiple testing, the Bonferroni method was used (*N*=10), resulting in a threshold of α=0.005. Significant effects of delay (for any latency relative to zero) were observed for vowel and syllable repetitions, vowel exchanges, vowel omissions, distortions, and onset disfluencies[4]. The following error-delay pairs showed a statistically significant (*p* < *0.005*) effect relative to the zero-delay condition: vowel repetitions (100, 150, 200, 250 ms), syllable repetitions (200, 250 ms), vowel exchanges (250 ms), distortions (150, 200, 250 ms), and onset disfluencies (100, 150, 200, 250 ms); these are marked with an asterisk in Figure 4.

3.3 Timing of error occurrence

We also analyzed the “arrival time” of errors within the 5 s production period. We hypothesized that increased delay would lead to earlier error arrival times. The raw average error arrival time generally decreased with increasing delay, from syllable number 12.6 at zero delay to syllable number 8.2 at 250 ms delay. A generalized linear mixed effects Poisson regression model (log link function) with random intercepts for subject and sequence and fixed effect of delay was fit to arrival times for the 845 classified error trials. A significant effect of delay (compared to zero delay) was found for latencies greater than 50ms. These results are summarized in Table 3.

[4] No examples of onset disfluency errors were observed at zero delay, and only one example was observed at 50 ms delay. These two conditions were combined as a baseline level for the factor delay in the model for this error type to allow model convergence. Confidence intervals on parameter estimates are large and caution is warranted in interpreting these results.

3.4 Speaking rate effects

A visual metronome signal was used to provide a pacing signal and reduce participants' tendency to slow their speech rate with increasing delay. We calculated the average inter-syllable duration in each trial, measured as the total duration of the transcribed speech on a given trial divided by the number of syllables transcribed. An overall trend to reduce rate with delay was observed, with average inter-syllable durations of ~220 ms at zero delay up to ~252 ms at 200 and 250 ms delays (across all trials and participants). To better understand this duration effect, we used a linear mixed effects model to estimate the effects of delay, run (a proxy for time in the experiment), and presence/absence of an error on inter-syllable durations. The model (summarized in an ANOVA table in Table 4; see Supplemental File 2 for the full set of model coefficients) demonstrated significant effects of run ($F(4)= 365.3$, $p<2.2\text{x}10^{-16}$) and delay ($F(5) = 207.7$, $p<2.2\text{x}10^{-16}$) as well as a run x delay interaction ($F(20) = 5.68$, $p<8.0\text{x}10^{-15}$). Figure 5 shows inter-syllable durations, averaged within participant, as a function of run number and delay. Durations increased (rate decreased) with increasing delay, while durations decreased (rate increased) with increasing run number (time in the experiment). The interaction indicates that the effect of delay on inter-syllable durations decreased as the experiment went on. From Figure 5 it is clear that participants produced longer inter-syllable durations in Run 1 (i.e., first 60 trials of the experiment) than in other runs. Pairwise comparisons (*P*-values adjusted using Tukey's method) over the variable *run* indicated that inter-syllable durations were significantly longer in Run 1 than in all other runs ($p<0.0001$); likewise, Run 2 durations were longer than Runs 3-5 ($p<0.0001$), and the difference between Run 3 and Run 5 was marginally significant ($p = 0.040$). Supplementary Figure 3 additionally shows an interaction plot (effect of delay for each run number) based on the fitted linear mixed effects model.

3.5 Temporal analysis of repetition errors

Syllable boundaries and vowel nuclei were marked for all syllable (N=48) and vowel (N=177) repetition errors. An analysis of the timing relationships between the two syllables involved in each repetition error was performed. Figure 2 shows an example syllable repetition error and illustrates the three temporal measures that were examined. Linear mixed effects models were fit to each measure. Because only 4 of these errors (3 vowel repetitions and 1 syllable repetition) occurred at 50 ms delay, and no errors occurred at zero delay, these were excluded from the analysis. The model estimates revealed significant effects of delay and error unit type (whether the syllable or vowel was repeated) for only the *overlap* measure (see Table 5; see Supplementary File 3 for *onset interval* and *peak interval* models). Overlap was higher for syllable repetition errors than for vowel repetition errors and increased with longer delay (though the only significant coefficient for delay was at 100 ms). No significant interactions were found. Figure 6 illustrates the distributions of the *overlap* measure from individual trial data as a function of delay and error unit type (syllable vs. vowel).

4 Discussion

In this study we sought to systematically measure the impact of delayed auditory feedback on speech sequencing errors using a highly controlled, paced syllable repetition task. We hypothesized that the use of a visual pacing signal alongside this stereotyped production task would allow participants to resist the tendency to reduce speech rate and highlight the auditory-motor interactions that drive disfluent speech. As discussed below, the visual metronome was partly effective; participants decreased their speaking rate with increasing delay to a considerably lesser extent than observed in previous studies. Using a simple but detailed transcription approach, our results provide a precise quantification of error patterns across a large number of trials in each speaker and across a range of delays (between 0 and 250 ms).

Our primary research objective was to characterize the types of serial order errors that increase under DAF. We defined discrete serial order error types based on extensive previous work on naturally occurring slips of the tongue (Fromkin, 1971; MacKay, 1970; Shattuck-Hufnagel, 1979; Vousden, Brown, & Harley, 2000), grouping errors into repetitions, anticipations, and exchanges. We additionally included errors involving substantially distorted productions and errors that approximated a “stutter” (Lee, 1951), in which repeated consonants were produced without a clear intervening vowel (labeled herein as *onset disfluencies*). Figures 3 and 4 demonstrate the patterns of errors made as a function of auditory feedback delay. A generalized linear mixed effects logistic regression model was fit to a binary dependent variable indicating the presence or absence of *any* error in a trial, revealing a strong effect of delay on error rate for delays of 100 – 250 ms. Models with the same structure were then fit to the presence or absence of individual error types. These results revealed that DAF elicits a *specific* pattern of speech errors in this task, as opposed to having a more general modulatory effect on all speech errors. These error patterns are further discussed below.

4.1 Errors predominantly involved vowel and syllable units

As can be seen in Figure 4, errors heard by transcribers as discrete sound substitutions relative to the target sequence (and not distortions) were most commonly *vowel repetitions* and, to a somewhat lesser extent, whole *syllable repetitions*. Vowel repetitions (e.g., ba-da-gu for target ba-di-gu) were significantly more frequent at delays of 100 ms or more (relative to zero delay), while syllable repetitions (e.g., ba-ba-di-gu for ba-di-gu) were significant at 200 and 250 ms delays only. Interestingly, consonant repetition errors were not significantly impacted by DAF. Furthermore, while consonants were the most common unit involved in *exchange errors* (e.g., ba-gi-du for target ba-di-gu) they were not significantly modulated by delay. Vowel exchanges did, on the other hand, significantly increase at the 250 ms latency only. Syllables exchanges and anticipations were extremely rare and were not modulated by delay. Furthermore, the relatively large number of distortion errors observed were driven almost entirely by vowels, with only a very small number (*N*=3) indicated due to atypical (or out of set / non-contextual) consonant productions.

Given the dominance of vowel and syllable unit errors, we can conclude that temporally mismatched feedback drives sound substitution errors for specific units in the speech plan, rather than affecting all units equally. The observed effects run counter to typically occurring speech errors, in which consonants are more commonly substituted than vowels (MacKay, 1970; Vousden et al., 2000). Indeed, consonant unit errors were more common in our results at zero delay but were largely unaffected by DAF. Why might this be the case? In the DAF paradigm, participants receive typical somatosensory feedback simultaneous with atypical auditory feedback. The onsets of each syllable in our paradigm (voiced stops) provide substantial tactile feedback that could have allowed speakers to sense that these sounds were being produced as expected. On the other hand, vowel productions (and syllable productions) provide relatively long duration, steady state auditory feedback that can be compared with expectations. Speakers monitor vowel productions, as evidenced by compensations to online shifts in formant frequencies (Cai et al., 2011; Lester-Smith et al., 2020; Niziolek & Guenther, 2013; Purcell & Munhall, 2006; Tourville, Reilly, & Guenther, 2008). Since auditory feedback alone was altered in the present study, it is plausible that the observed error patterns highlight differences in the relative importance of different feedback modalities for different segment types. We cannot rule out, however, that the differential effects are related to CV structure, impacting vowels due to where they appear within the syllable. Testing with VC syllables, for example, might result in a different pattern of errors; we are exploring this possibility in a larger dataset currently being collected in the lab that uses more diverse speech materials.

It should be noted that errors linked to consonant productions appear in two of the other error types examined: onset disfluencies and vowel omissions. These were not classified as consonant repetition errors (Table 1) because they cause a general change in the CVCVCV sequence and could be considered atypical, disfluent speech rather than fluent substitution errors. These two error types showed a general trend for increasing error probability as a function of delay time. 45 errors were classified as onset disfluencies and 66 errors were classified as vowel omissions. Vowel omission errors reached our threshold for significance only at 250 ms delay. No onset disfluency errors were observed at zero delay, which caused technical problems with model estimation. We combined the 0 and 50 ms condition and re-estimated the model, which then yielded significant increases for 100, 150, and 250 ms delays. Compared with other studies using different materials, these disfluency errors were relatively uncommon. It is possible that using nonword sequences that required maintenance in phonological working memory may have biased errors toward discrete substitutions and away from such disfluencies. Furthermore, it may have been somewhat difficult for listeners to clearly distinguish onset disfluencies (e.g., b-ba di gu) from consonant and/or syllable repetition errors (e.g., ba-ba-di-gu), which are similar but with the latter including a clear vowel production between consonant articulations. As with all listening-based error classification, certain ambiguities are unavoidable. We note, however, that to be included in the final data set, two listeners had to agree on the same error type (i.e., make the same call in these relatively ambiguous cases). Further examination of the detailed acoustics in both error types may help shed light on the precise nature of these induced disfluencies.

4.2 Substitution errors were dominated by repetitions

Figure 4 also demonstrates that *repetition* errors were far more commonly induced (at least as the initial error in a trial) by auditory feedback delays than anticipation or exchange errors. Again, this runs counter to results from naturally occurring slips of the tongue. For example, Vousden et al. (2000) analyzed a corpus of 2289 speech errors and found that 35.1% involved sound anticipations while 26% involved repetitions or perseverations. Stemberger (1989) also found that adult speech errors tended toward anticipations, accounting for 60% of combined anticipation and perseveration errors. Thus, DAF drives a distinct shift in the mode of errors produced, such that speech output becomes “focused on the past” (Dell, Burger, & Svec, 1997). This makes sense, since DAF brings feedback from past productions into perceptual time windows typically associated with the present output. This atypical feedback seems to directly drive repetitions under DAF. Possible mechanistic explanations for how this might occur in a model of speech production are discussed below.

A general shift from anticipatory toward perseverative errors has been observed in patients with aphasia, in children and adults producing unfamiliar speech sequences, and in individuals speaking more rapidly than normal (Schwartz, Saffran, Bloch, & Dell, 1994). Dell and colleagues (Dell et al., 1997; Martin & Dell, 2004) describe perseveration (of which repetition in one example) patterns in terms of residual activation of previous items in a phonological output buffer. A similar account can be offered by the GODIVA model (Bohland et al., 2010) and is described briefly below. In Dell and colleagues’ account, in patients with aphasia or other neurological damage, residual activation that drives repetition may be due to incremental learning that strengthens previously used connections or by a selection bias toward common or intact, undamaged representations. Interestingly, DAF has been proposed as a behavioral model for primary progressive aphasia (PPA) since its effects in healthy controls mimic some aspects of the disorder (Maruta et al., 2014). In the theoretical explanation from Dell and colleagues, in unfamiliar sequences, the shift to repetition errors effect may be driven by reduced overall activation of the target units. Finally, in rapid speech, the reduced time between activation of subsequent units provides less decay time for the previous unit and more opportunity for it to be reselected in error. In the present study, participants were presented with unfamiliar nonword sequences, but the closed set of six phonemes was repeatedly used, so familiarity grew over time. Participants had to produce these items rapidly (at 5 Hz), so relative novelty and rapid rate may have contributed to the observed dominance of repetition errors. However, Dell et al.'s model cannot explain why increased feedback latency drove increased dominance of these errors. Later, we discuss possible mechanisms for this effect based on the idea of residual activation.

4.3 Production of repeated sounds largely overlapped with relevant feedback

We further examined the vowel and syllable repetition errors that were elicited under DAF using an approach similar to that developed by Davis and Brajot (2019). We found that the amount of *overlap* between (i) the production of the second syllable involved in the repetition error and (ii) the auditory feedback signal from the first syllable involved in the repetition error (see Figure 2) was quite high (median values of 92.6% for syllable repetitions and 78.7% for vowel repetitions)[5], meaning that repetition errors were produced largely while the feedback from the previous syllable was present. The significant effect of error type indicated that overlap was greater for syllable repetitions (e.g., ba-ba) than for vowel repetitions (e.g., ba-da). No significant effects were observed for the other two temporal measures examined (*onset interval* or *peak interval*).

The results for the *overlap* measure are consistent with the possibility that auditory feedback signals themselves could, at least sometimes, drive the repeated productions of the same sounds. That is, in most cases, the participant was producing the repeated sound unit (whole syllable or vowel only) while hearing their feedback from the previous production. For auditory feedback from a previously produced syllable to influence the *production* of a syllable repetition error, it must arrive prior to the onset of production of the next syllable. However, if feedback from the first syllable arrived slightly later, it might only have sufficient time to influence the vowel produced in the subsequent CV syllable. Thus, the fact that overlap was higher when a full syllable repetition error occurred than when a vowel only repetition occurred supports this view. Further, while there was not a significant effect of error type for the *onset interval* measure (*p=0.10*), average values were larger for syllable repetition errors (onset of syllable feedback occurred on average ~36 ms before onset of the repeated syllable) than for vowel-only repetition errors (onset of syllable feedback occurred on average ~12ms before onset of the

[5] Note that the same syllable-level overlap measure was calculated for both syllable and vowel repetitions. Thus, any difference cannot be attributed to different units involved in the calculations themselves.

syllable containing the vowel repetition). These ideas are further discussed below in relation to existing models and possible computational mechanisms.

4.4 Errors occurred more often and arrived earlier for longer feedback latencies

A key manipulation in this study that differed from some others (though see, e.g., Corey & Cuddapah, 2008; Stuart, Kalinowski, Rastatter, & Lynch, 2002) was the systematic manipulation of delay (from near-zero to 250 ms). Like most previous studies, we found that errors were most prevalent at delays of 200-250 ms (in our study, errors were slightly more common at 250 ms than 200 ms delay), but we observed graded patterns such that errors occurred, but with lower probability, for lower delays (statistically significant for delays of 100 ms or greater; see Figure 4). An open question for future research is the extent to which these error probability functions (vs. delay) are scaled or even reshaped by other task parameters (i.e., metronome, feedback intensity, stimulus construction).

A secondary research question asked if errors arrived *earlier* in trials with longer delays than in trials with shorter delays. Speaking under DAF often elicits a subjective feeling of a struggle to overcome interference, which seems to accumulate over time. This idea was supported by our analysis, as the first error in an errorful trial occurred more than 4 syllables earlier, on average, for productions with a 250 ms delay compared to productions with zero delay. The mechanistic cause of this observed difference is not obvious. It is possible that a neural signal indexing “auditory error” accumulates over time, probabilistically increasing the chance of making a serial order error, and that longer delays give rise to additional auditory error for each syllable produced / heard. On the other hand, this accumulation effect may not occur at the level of the speech controller, but rather at a cognitive level due to, for example, increased attention and awareness of the mismatched auditory signal. Future work should attempt to clarify these potential effects.

4.5 Effects on speaking rate

In the early study by Fairbanks (Fairbanks, 1955), participants slowed the rate at which they read a sentence from the Rainbow Passage by an average of 70.4% when speaking with a 200ms delay compared to when speaking with zero delay. Davis and Brajot (2019) found that average syllable durations (from the first two sentences of the Rainbow Passage) were, on average, 30-40% longer when produced under a 200 – 300 ms delay than when produced with zero (minimal) delay. These authors cast these speaking rate reductions as a “partial compensation” for the delay (cf. partial compensation for pitch or formant-shifted feedback) and showed that the extent of compensation (duration increase divided by delay latency) decreased gradually with increasing delay.

In our study, we used a visual metronome as a pacing signal with a target inter-syllable duration of 200 ms (5 Hz). The estimated marginal mean inter-syllable duration increased from 220 ms to 252 ms from the zero delay to 250 ms delay conditions (the median increased from 217 ms to 236 ms). Thus, on average, speakers slowed their speech by only ~14.5% at the longest delay. Casting these values as “partial compensations” we also see considerably smaller compensations (~15.5% at 200 ms) than in the previously reported study (Davis & Brajot, 2019). These substantial differences in rate reduction can likely be attributed to the use of a visual pacing signal in the present study. It is also possible that the use of simple, stereotyped CVCVCV stimuli, combined with the visual metronome was critical to the decreased rate reductions observed here; ongoing work in our lab, however, suggests that visual pacing signals are also useful in maintaining typical production rates for more complex nonwords and sentences.

We also observed (see Figure 5 and Supplemental Figure 1) a significant main effect of run order on inter-syllable durations as well as a significant interaction between run and delay. In particular, the first run of each session was characterized by substantially slower productions (mean 265 ms vs. 227 ms in the final run). Pairwise comparisons across runs indicated that participants significantly reduced rate from Run 1 to Run 2, and then again from Run 2 to Run 3; Runs 3 through 5 resulted in very similar speaking rates, although there was a marginally significant difference between Runs 3 and 5 ($p=0.04$). Because delay intervals were pseudorandomized, participants were not able to adapt to a specific delay. Therefore, we attribute the reduced slowing across runs to general increased comfort with the task and an increased ability to make use of the visual metronome signal over time. There was one participant (labeled S9 in Figure 5) who was consistently unable to produce the stimuli at a rate that approached the metronome. This participant made errors on a substantial number of trials (see Figure 3) but not as often as some other participants. It is possible that this individual exhibited a speed-accuracy tradeoff, reducing rate to, in turn, reduce the overall number of errors. However, in general, no relationship was apparent between participants’ mean duration and error rate.

4.6 Toward a model-based account of speech serial order errors under DAF

In the earliest discussions around DAF, researchers focused on explanatory accounts based on simple feedback controllers. Black (Black, 1951) proposed that speakers attempt to rely on the delayed feedback signal for error monitoring and thus slow down as a compensatory mechanism. Levelt and colleagues (Levelt, 1989; Levelt, 1983; Levelt et al., 1999) proposed the *perceptual loop hypothesis*, whereby speakers make use of the speech comprehension system to monitor errors. These processes are proposed to operate on both inner speech and external speech. Under DAF, the inner speech signal should remain error-free, but the outer (external) speech loop would sense errors, triggering an interruption and attempts to self-repair. These processes are relatively slow (estimated to require ~350-

400 ms; Gauvin & Hartsuiker, 2020) and are unlikely to explain the specific error patterns observed here.

The DIVA model (Guenther, 2016) proposes that incoming auditory feedback is compared with auditory target regions, which are read out when sound units (typically syllables) are activated in the "Speech Sound Map" (SSM) for production. The mismatch between these targets and feedback induced by DAF would result in auditory error signals, which, in DIVA, would drive changes in productions (via the feedback control subsystem) in an attempt to reduce error. These corrective signals are based on the difference in the delayed feedback signal (perhaps related to the *previous syllable*) and the auditory expectations for the *current syllable*. This mechanism would be expected to elicit atypical productions that might correspond to the "distortion errors" obtained in the present study. Under the paced speaking conditions used here, the opportunities for such corrective actions, however, are somewhat limited as feedback-based control requires time to register and process the incoming auditory signals, which were designed to be relatively short-lasting, and compute and enact compensatory motor actions.

The feedback control system in the DIVA model (see Guenther, 2016 for a detailed description of the proposed mechanisms) cannot directly account for the pattern of discrete sound errors observed here and in other DAF studies[6]. GODIVA (Bohland et al., 2010), however, adds proposed representations of speech sequences and their serial order (a phonological output buffer), providing a potential basis for explaining these results. In GODIVA, a syllable sequence is represented by parallel competitive queuing representations (Bullock & Rhodes, 2003; Grossberg, 1978a, 1978b; Houghton & Hartley, 1995) for (i) the abstract syllable frame (e.g., CV), and (ii) the phonemes that make up the sequence. A simplified schematic is provided in Figure 7. Serial order is represented by the relative activation levels of the representative units, and errors occur when one unit's activity erroneously exceeds the activity of the target unit at the time of response selection. As suggested by Martin and Dell (2004), any model of serial order requires (i) a mechanism to turn off past sound units, (ii) a mechanism to activate the present sound unit, and (iii) a mechanism to prime future sound units. The first mechanism is naturally linked to sound repetition errors. In GODIVA, constituent phonemic representations (which are distinct for onsets and vowels) are suppressed when an appropriate syllable-sized motor program is selected for production in the Speech Sound Map, in turn allowing the next items to activate the next syllable, and so on. As syllables are iteratively produced, predictive

[6] However, note that Civier et al. (2010) proposed the addition of an "Excessive error detector" within a "Monitoring subsystem" to the DIVA model, which might drive resets to the motor program in people who stutter. This subsystem was also suggested as a possible mechanism for "resets" that occur under DAF (Chesters, Baghai-Ravary, & Möttönen, 2015) but has not been further elaborated.

completion signals suppress those SSM representations, allowing activation and production of the next syllable.

In principle, auditory feedback has no clear mechanism in the model to influence the residual activation of these suppressed phonological units. Given typical speaking rates, it is not feasible that suppression (and in turn selection and initiation of the next sounds) can wait for external auditory feedback signals (Houde & Nagarajan, 2011). Here we propose two possible computational mechanisms that could be added to the GODIVA model (or similar models) to begin to account for the observed effects. Further experimental and computational efforts are required to test the feasibility and fit of such accounts to these and other data.

First, while auditory feedback cannot be relied upon for response suppression due to its sluggishness, the consistent, large auditory error signals caused by DAF could give rise to a general *conflict signal* within the production system that might dampen control signals due to uncertainty and speaker hesitancy. Weakened control signals might then lead to reduced response suppression and additional residual activation of previously produced sounds, increasing the probability of repetition errors. The reduced suppression might act at the phonological level (onset and vowel cells in Figure 7) or at the syllable level – both of these representations require strong inhibitory suppression signals. Under this account, the mismatched auditory feedback received at any specific point in time is unimportant, but rather the overall consistent errors elicited under DAF would drive a propensity to repeat due to a failure to suppress previously chosen items. Because we did not find evidence for an overall increase in consonant repetition errors, however, the auditory error signals would need to be preferentially linked to vowels and syllables, whereas perhaps somatosensory error signals (not affected by DAF) would be linked to the representations of onset consonants used in this task. This potential mechanism could explain the effect of error arrival time as the conflict signal would be expected to accumulate over time, with overall conflict increasing more rapidly for longer delays that result in higher degrees of mismatch with targets.

A second possibility is that the *specific* auditory feedback signals contribute in real-time to the elicited errors. Speech perception tasks activate speech motor representations (D’Ausilio et al., 2009; Du, Buchsbaum, Grady, & Alain, 2014; Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Wilson, Saygin, Sereno, & Iacoboni, 2004), and sound inputs that match a to-be-produced word facilitate responses and reduce reaction times (Galantucci, Fowler, & Goldstein, 2009; Meyer, 1990, 1991). In speech with typical feedback, the auditory input matches expectations, and auditory responses are suppressed compared to passive listening (Heinks-Maldonado, Nagarajan, & Houde, 2006; Houde, Nagarajan, Sekihara, & Merzenich, 2002). Under DAF, however, these responses may not be similarly suppressed and may instead act more like externally generated speech inputs when sound feedback fails to occur within the expected temporal windows. If true, these activations could “prime” speech outputs by providing additional activation to their corresponding representations in either the phonological planning buffer (for vowels) or SSM (for syllables). This additional activation would then directly increase the probability of repetition errors involving the activated sounds. Following the assumption that sound units with longer acoustic durations (e.g., vowels and syllables) are more likely to prime speech output units in this manner, we would then expect the observed preponderance of errors to involve these units.

We conducted a detailed exploratory timing analysis of repetition errors involving vowels and whole syllables (see Figure 2 for illustration of the measures examined). We found evidence to support the idea that, in the present task, participants typically produced a repeated vowel or syllable largely overlapping with the auditory feedback of the previous syllable, and the extent of overlap was significantly greater for whole syllable repetition errors than for vowel-only repetitions. Thus, the relevant auditory feedback that *could* induce an error is often present at the time the error is initiated. Figure 2 provides one specific example of a syllable repetition error (participant repeated the syllable /du/). Here, the feedback from the first production of /du/ begins prior to the onset of production of the repetition and has 100% *overlap*. We propose here that this unexpected, unsuppressed feedback from the first /du/ drives extra activation of the /du/ plan, facilitating a syllable repetition error. The observed difference in *overlap* for syllable vs. vowel repetition errors makes sense; for example, if the feedback for the first production arrived slightly after the onset of production of the next syllable, it might still influence a vowel repetition but clearly could not be the source of a syllable repetition error.

These findings are consistent with the possibility that the auditory inputs from previously produced sounds might influence their repeated productions. However, this action would need to be very rapid – much faster, for example, than what is seen for processing auditory error and issuing corrective actions in auditory perturbation studies (Tourville et al., 2008). Latencies of the cortical frequency following response to aurally presented speech sounds measured using intracranial stereoelectroencephalography (sEEG) in humans have been reported to be as short as ~17 ms (Gnanateja et al., 2021). Thus, it appears possible that, at least in many cases, incoming auditory information has the potential to influence planned speech quite rapidly. Further work should investigate the timing of these possible mechanisms as well as their ability to account for repetitions and other disfluencies observed in persistent developmental stuttering, where sensory-motor timing may be atypical (Etchell, Johnson, & Sowman, 2015; Sares, Deroche, Shiller, & Gracco, 2018, 2019).

4.7 Limitations and Future Work

While the current study provides an in-depth, interpretable analysis of speech errors elicited by delayed auditory feedback in a highly controlled task, there are a number of limitations that motivate ongoing and future work. Most obviously, the present results are based on “lab speech” – using a contrived atypical, rhythmic speaking task that may limit generalizability to other speaking contexts. In particular, the use of CVCVCV sequences only here presents two limitations. First, we cannot assume that the effects of the visual metronome or the overall effects of DAF are representative of their effects on other more complex or linguistic stimuli. Second, these stimuli conflate position within the syllable with the consonant-vowel distinction; thus, it is important to rule out that differences in consonant vs. vowel errors observed here are not due to syllable position effects. Ongoing work in our laboratory seeks to better understand DAF-induced effects as a function of speech planning load and utterance complexity and will also directly compare errors during nonword syllable sequence production and sentence production.

In our analyses, only the *first error* produced in a trial was analyzed. Speakers often made additional errors, which may be informative in understanding the overall effects of DAF; however, because the sometimes-conscious detection of the first error appeared to drive diverse and complex compensatory strategies, we limited current analysis to what was most approachable. The classification of even these first errors also had certain ambiguities. For example, a repetition error (e.g., ba-da-gu for target ba-di-gu) could also, in principle, be an *anticipation* error as the “/a/” vowel must be produced again later for the *next* production of the entire sequence, which was repeated multiple times per trial. Our approach here was the parsimonious one – simply classifying errors based on minimum transposition distance. It is also important to note that, because our results categorize each trial into only a single category based on the first error, they should not be taken to indicate the probability that that type of error occurred *at all* in each trial. Indeed, all of the presented error rates (Figure 4) likely undershoot the rates we would have obtained if multiple errors could be unambiguously classified across the full utterances.

Error analyses also did not consider detailed acoustics of the produced sounds. It is extremely likely that sub-phonemic articulatory errors (or distorted productions) were present in trials marked as “error free” or as discrete sequencing errors. The evaluation of speech errors by listening to audio recordings may be influenced by perceptual biases, particularly resulting in judgments of “phonemic” errors that could have a sub-phonemic basis (Cutler, 1981; Pouplier & Hardcastle, 2005). This idea is supported by an electromagnetic articulography (EMA) analysis, which found more variable articulations for syllables produced under DAF (200 ms delay) compared to zero delay, even when those productions were judged as typical by listeners (Cler et al., 2017). Detailed examination of acoustics (for example vowel formant frequencies) might provide a more rigorous analysis of some errors, especially distortion errors, but is extremely laborious and subject to human error for data of this volume.

Because our experiment involved only delayed *speech* signals, it is also not possible to determine definitively how much the actual acoustic content of the feedback (as opposed to the timing of more general sound inputs) impacted the present results. Previously, a non-speech noise signal, matched to the speech envelope and delayed, was found to cause similar disruption of ongoing speech production (Howell & Archer, 1984). This disruption, however, was measured using a rate variable and did not consider whether or not the errors encountered in delayed speech versus delayed non-speech noise were similar or different. This is a question that warrants further investigation; determining precisely what *matters* in the feedback signal will greatly help to constrain computational models.

Finally, our sample (effectively N=14 participants) is limited and relatively homogenous. Given that the effects of DAF are highly variable, it is worthwhile to pursue larger sample sizes such as those used by Chon et al. (2013). Gender differences, for example, have been previously studied with somewhat inconclusive results (Corey & Cuddapah, 2008; Stuart & Kalinowski, 2015). Larger samples, while clearly better for understanding individual variability – which was not one of the goals of this study – typically make the tradeoff of reducing the number of data points *within-subject*. Here we chose to gather extensive samples from each individual speaker (productions of 360 repeated CVCVCV sequences in most participants), which allowed us to assess effects of delay while modeling inter-subject variability using random intercepts in mixed-effects models. It is evident from Figure 3 that our cohort did have substantial variability. For example, two participants (S7 and S13) made relatively few errors and two others (S9 and S12) were also more modestly impacted than others. 4 out of 14 (~28%) low-responding participants is comparable to the proportion reported by Chon et al. (2013) using spontaneous speech, who showed 18/62 (~29%) or 33/62 (~53%) subjects falling within a 'low' responder group when clustering subjects into two or three groups. Ongoing work in our lab is using a much larger sample size and more varied speaking materials to assess potential sources and explanations for such variation.

4.8 Summary and Conclusions

In this study we developed a highly structured and controlled behavioral task designed to elicit speech errors under delayed auditory feedback. Our primary objective was to characterize the types of serial order errors that occur when auditory feedback is misaligned with expectations, while speech proceeds at a rapid rate. We showed that a selective set of sequencing errors, most notably including vowel and whole syllable repetitions, increased with increasing feedback delays. We also demonstrated that errors tended to arrive earlier in a trial as feedback delays increased, suggesting a possible accumulation of error over time. Our protocol used a visual metronome signal to encourage participants to trade fluent production for increased speaking rate. This pacing signal was partly effective; though participants still slowed their speech somewhat with increasing delays, the extent of rate reduction was less than in previous studies, which likely resulted in the elicitation of additional errors. Finally, we introduced two potential mechanistic extensions of the GODIVA model of speech planning and production, which make different predictions, but each offer the potential to at least partially account for the patterns of speech errors observed under DAF.

5 Acknowledgments

This work was supported by NSF Awards SMA 0835976 (PI Shinn-Cunningham) and BCS 2029245 (PI Bohland) as well as the Boston University Undergraduate Research Opportunities Program (UROP). We additionally thank Cara Stepp, Vincent Pennetti, Emily Hacker, and Ashley Petitjean for their assistance in various stages of this work.

6 References

Ambrose, N. G., & Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research*, *42*(4), 895–909. https://doi.org/10.1044/jslhr.4204.895

ATKINSON, C. J. (1953). Adaptation to delayed side-tone. *The Journal of Speech and Hearing Disorders*, *18*(4), 386–391. https://doi.org/10.1044/jshd.1804.386

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Behroozmand, R., Korzyukov, O., Sattler, L., & Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control. *The Journal of the Acoustical Society of America*, *132*(4), 2468. https://doi.org/10.1121/1.4746984

Black, J. W. (1951). The effect of delayed side-tone upon vocal rate and intensity. *The Journal of Speech Disorders*, *16*(1), 56–60. https://doi.org/10.1044/jshd.1601.56

Boersma, P., & Weenink, D. (2001). *Praat, a system for doing phonetics by computer*. Retrieved from https://scholar.google.com/scholar?hl=en&q=praat&btnG=&as_sdt=1%2C22&as_sdtp=#0

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*(7).

https://doi.org/10.1162/jocn.2009.21306

Bullock, D., & Rhodes, B. (2003). Competitive queuing for planning and serial performance (M. A. Arbib, Ed.). *The Handbook of Brain Theory and Neural Networks*, pp. 241–244. Cambridge, MA: MIT Press.

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, *103*(6), 3153–3161. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9637026

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. *Journal of Neuroscience*, *31*(45), 16483–16490. https://doi.org/10.1523/JNEUROSCI.3653-11.2011

Chesters, J., Baghai-Ravary, L., & Möttönen, R. (2015). The effects of delayed auditory and visual feedback on speech production. *The Journal of the Acoustical Society of America*, *137*(2), 873–883. https://doi.org/10.1121/1.4906266

Chon, H. C., Jo Kraft, S., 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*(2), 489–504. https://doi.org/10.1044/1092-4388(2012/11-0303)

Civier, O., Tasko, S. M., & Guenther, F. H. (2010). Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *Journal of Fluency Disorders*, *35*(3), 246–279. https://doi.org/10.1016/j.jfludis.2010.05.002

Cler, G. J., Lee, J. C., Mittelman, T., Stepp, C. E., & Bohland, J. W. (2017). Kinematic analysis of speech sound sequencing errors induced by delayed auditory feedback. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1695–1711. https://doi.org/10.1044/2017_JSLHR-S-16-0234

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

Corretge, R. (2020). *Praat Vocal Toolkit*. Retrieved from http://www.praatvocaltoolkit.com

Cutler, A. (1981). The reliability of speech error data. *Linguistics*, *19*(7–8), 561–582. https://doi.org/10.1515/LING.1981.19.7-8.561/HTML

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385.

https://doi.org/10.1016/j.cub.2009.01.017

Davis, S. N., & Brajot, F.-X. (2019). Partial compensation to delayed auditory feedback: An analysis of syllable duration. *The Journal of the Acoustical Society of America*, *145*(6), 3531–3540. https://doi.org/10.1121/1.5111758

De Andrade, C. R. F., & Juste, F. S. (2011). Systematic review of delayed auditory feedback effectiveness for stuttering reduction. *Jornal Da Sociedade Brasileira de Fonoaudiologia*, *23*(2), 187–191. https://doi.org/10.1590/S2179-64912011000200018

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: a functional analysis and a model. *Psychological Review*, *104*(1), 123–147. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9009882

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(19), 7126–7131. https://doi.org/10.1073/pnas.1318738111

Etchell, A. C., Johnson, B. W., & Sowman, P. F. (2015). Beta oscillations, timing, and stuttering. *Frontiers in Human Neuroscience*, *8*(JAN), 1036. https://doi.org/10.3389/fnhum.2014.01036

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*(2), 399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x

Fairbanks, G. (1955). Selective vocal effects of delayed auditory feedback. *Journal of Speech & Hearing Disorders*, *20*(4), 333–346. Retrieved from http://doi.apa.org/?uid=1956-07109-001

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

Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52.

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, and Psychophysics*, *71*(5), 1138–1149. https://doi.org/10.3758/APP.71.5.1138

Gauvin, H. S., & Hartsuiker, R. J. (2020). Towards a new model of verbal monitoring. *Journal of Cognition*, *3*(1), 1–37. https://doi.org/10.5334/JOC.81/METRICS/

Gnanateja, G. N., Rupp, K., Llanos, F., Remick, M., Pernia, M., Sadagopan, S., … Chandrasekaran, B. (2021). Frequency-Following Responses to Speech Sounds Are Highly Conserved across Species and Contain Cortical Contributions. *ENeuro*, *8*(6), ENEURO.0451-21.2021.

https://doi.org/10.1523/ENEURO.0451-21.2021

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. *Progress in Theoretical Biology*, *5*, 233–374.

Grossberg, S. (1978b). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, *17*(3), 199–219. https://doi.org/10.1016/0022-2496(78)90016-0

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A Theoretical Investigation of Reference Frames for the Planning of Speech Movements. *Psychological Review*, *105*(4), 611–633. https://doi.org/10.1037/0033-295X.105.4.611-633

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport*, *17*(13), 1375–1379. https://doi.org/10.1097/01.wnr.0000233102.43526.e9

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2011.01.019

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

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience*, *14*(8), 1125–1138. https://doi.org/10.1162/089892902760807140

Houghton, G., & Hartley, T. (1995). Parallel models of serial behavior: Lashley revisited. *Psyche*, *2*, 2–25.

Howell, P., & Archer, A. (1984). Susceptibility to the effects of delayed auditory feedback. *Perception & Psychophysics*, *36*(3), 296–302. https://doi.org/10.3758/BF03206371

Kalinowski, J., Armson, J., Stuart, A., & Gracco, V. L. (1993). Effects of alterations in auditory feedback and speech rate on stuttering frequency. *Language and Speech*, *36*(1), 1–16. https://doi.org/10.1177/002383099303600101

Kalinowski, J., Stuart, A., Sark, S., & Armson, J. (1996). Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language and Communication Disorders*, *31*(3), 259–269. https://doi.org/10.3109/13682829609033157

Kim, K. S., Wang, H., & Max, L. (2020). It's about time: Minimizing hardware and software latencies in speech research with real-time auditory feedback. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2522–2534. https://doi.org/10.1044/2020_JSLHR-19-00419

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models . *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lee, B. S. (1950). Effects of Delayed Speech Feedback. *Journal of the Acoustical Society of America*, *22*(6), 824–826. https://doi.org/10.1121/1.1906696

Lee, B. S. (1951). Artificial stutter. *Journal of Speech & Hearing Disorders*, *16*(1), 53–55. https://doi.org/10.1044/JSHD.1601.53

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(1), 3.

Lester-Smith, R. A., Daliri, A., Enos, N., Abur, D., Lupiani, A. A., Letcher, S., & Stepp, C. E. (2020). The relation of articulatory and vocal auditory–motor control in typical speakers. *Journal of Speech, Language, and Hearing Research*, *63*(11), 3628–3642. https://doi.org/10.1044/2020_JSLHR-20-00192

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 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. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

MacKay, D. G. (1970). Spoonerisms: the structure of errors in the serial order of speech. *Neuropsychologia*, *8*(3), 323–350.

Martin, N., & Dell, G. S. (2004). Perseverations and anticipations in aphasia: primed intrusions from the past and future. *Seminars in Speech and Language*, *25*(4), 349–362. https://doi.org/10.1055/s-2004-837247

Maruta, C., Makhmood, S., Downey, L. E., Golden, H. L., Fletcher, P. D., Witoonpanich, P., … Warren, J. D. (2014). Delayed auditory feedback simulates features of nonfluent primary progressive aphasia. *Journal of the Neurological Sciences*, *347*(1–2), 345–348. https://doi.org/10.1016/j.jns.2014.09.039

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The Essential Role of Premotor Cortex in Speech Perception. *Current Biology*, *17*(19), 1692–1696. https://doi.org/10.1016/j.cub.2007.08.064

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*(5), 524–545. https://doi.org/10.1016/0749-596X(90)90050-A

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, *30*(1), 69–89.

https://doi.org/10.1016/0749-596X(91)90011-8

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience*, *33*(29), 12090–12098. https://doi.org/10.1523/JNEUROSCI.1008-13.2013

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1051–1059. https://doi.org/10.1044/1092-4388(2010/10-0162)

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, *77*(2), 97–132.

Pouplier, M., & Hardcastle, W. (2005). A re-evaluation of the nature of speech errors in normal and disordered speakers. *Phonetica*, *62*(2–4), 227–243. https://doi.org/10.1159/000090100

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *NA Report NA2009/06*, 39. Retrieved from https://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

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*(4), 2288. https://doi.org/10.1121/1.2173514

Sares, A. G., Deroche, M. L. D., Shiller, D. M., & Gracco, V. L. (2018). Timing variability of sensorimotor integration during vocalization in individuals who stutter. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-34517-1

Sares, A. G., Deroche, M. L. D., Shiller, D. M., & Gracco, V. L. (2019). Adults who stutter and metronome synchronization: evidence for a nonspeech timing deficit. *Annals of the New York Academy of Sciences*, *1449*(1), 56–69. https://doi.org/10.1111/nyas.14117

Schwartz, M. F., Saffran, E. M., Bloch, D. E., & Dell, G. S. (1994). Disordered speech production in aphasic and normal speakers. *Brain and Language*, *47*(1), 52–88. https://doi.org/10.1006/brln.1994.1042

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W.E. Cooper & E.C.T. Walker (Eds.), *Sentence processing: Studies dedicated to Merrill Garrett* (pp. 295–342). Hillsdale, NJ: Erlbaum.

Soderberg, G. A. (1969). Delayed auditory feedback and the speech of stutterers: a review of studies. *The Journal of Speech and Hearing Disorders*, Vol. 34, pp. 20–29. https://doi.org/10.1044/jshd.3401.20

Stemberger, J. P. (1989). Speech errors in early child language production. *Journal of Memory and Language*, *28*(2), 164–188. https://doi.org/10.1016/0749-596X(89)90042-9

Stuart, A., & Kalinowski, J. (2015). Effect of delayed auditory feedback, speech rate, and sex on speech production. *Perceptual and Motor Skills*, *120*(3), 747–765. https://doi.org/10.2466/23.25.PMS.120v17x2

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*(5), 2237. https://doi.org/10.1121/1.1466868

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, *39*(3), 1429–1443. https://doi.org/10.1016/j.neuroimage.2007.09.054

Vousden, J. I., Brown, G. D., & Harley, T. a. (2000). Serial control of phonology in speech production: a hierarchical model. *Cognitive Psychology*, *41*(2), 101–175. https://doi.org/10.1006/cogp.2000.0739

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

Yates, A. J. A. (1963). Delayed auditory feedback. *Psychological Bulletin*, *60*(3), 213–232. https://doi.org/10.1037/h0044155

7 Figures

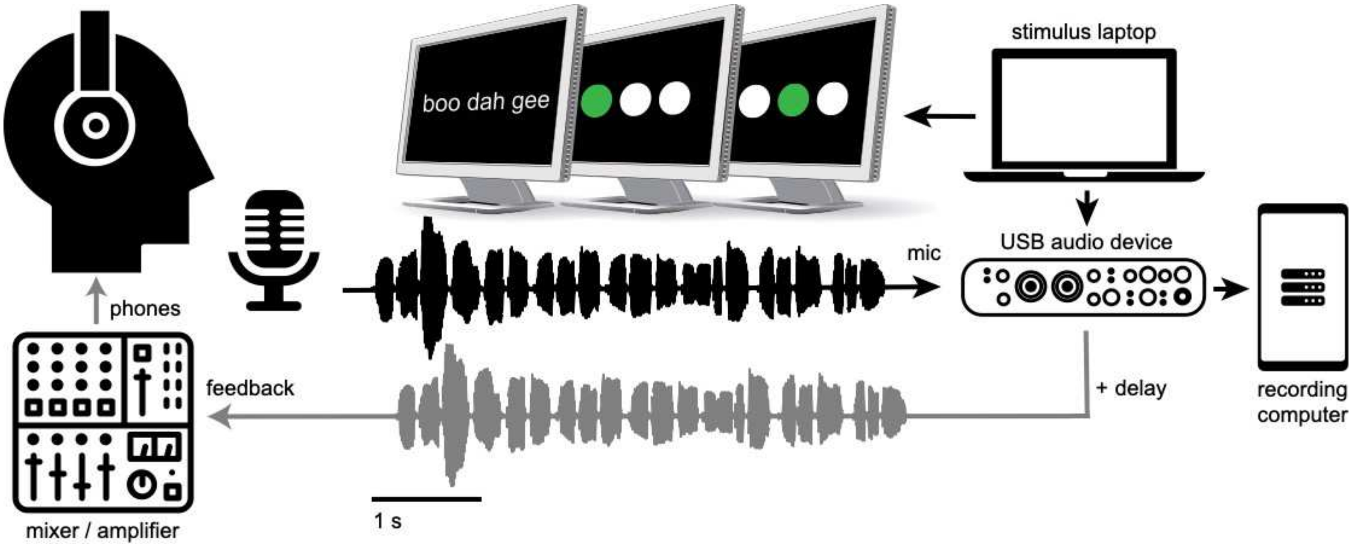

Figure 1. Illustration of experimental protocol. Participants viewed stimuli presented orthographically on a screen and repeatedly produced the presented sequence, paced to a visual metronome (three circles) signal. Signals were recorded using a head-worn condenser microphone (Shure WH30XLR) attached to a USB audio device (M-Audio FastTrack Ultra). This device transmitted a delayed signal, which was amplified and fed back to the participant over circumaural headphones (Sennheiser HD280 Pro).

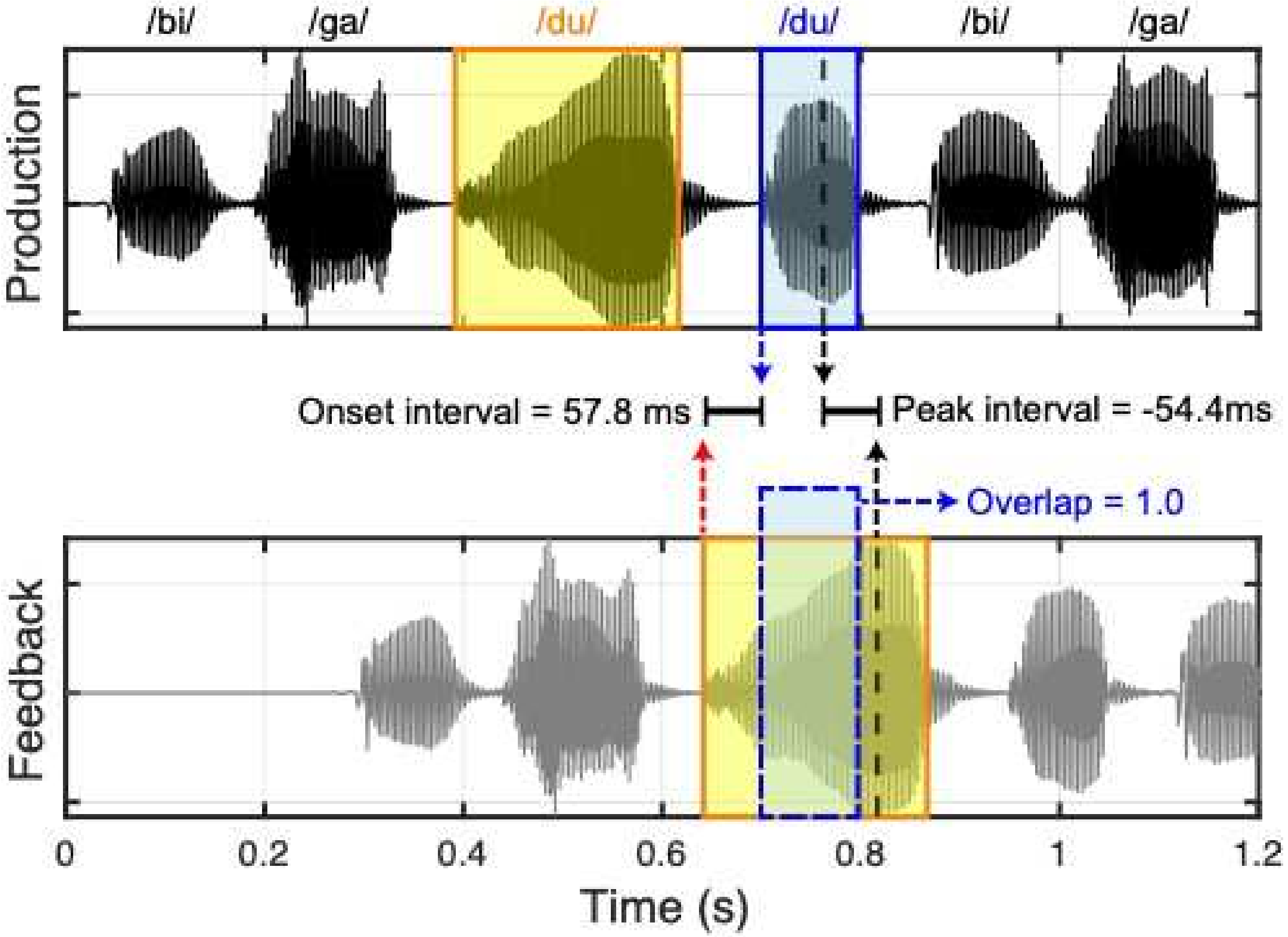

Figure 2. Illustration of timing of a syllable repetition error. *Top*: Production (microphone) signal showing the first 6 syllables produced in a sample error trial (target sequence bi-ga-du) by one participant. The participant produced the syllable /du/ as expected (orange box), then repeated that syllable (blue box). *Bottom*: Feedback (headphone) signal showing delayed feedback, time aligned to the production signal. The production of the repeated syllable /du/ (indicated by the dashed blue box) occurred completely within the duration of the feedback for the first production of /du/ (indicated by the orange box), resulting in an *overlap* value of 1.0 for this trial. The *onset interval* indicates that the onset of the repeated /du/ syllable occurred ~57.8 ms after the onset of feedback for the first /du/ production. The *peak interval* indicates that the sonority peak for the repeated /du/ syllable occurred ~54.4 ms *before* the sonority peak of the first /du/ production occurred in the feedback channel.

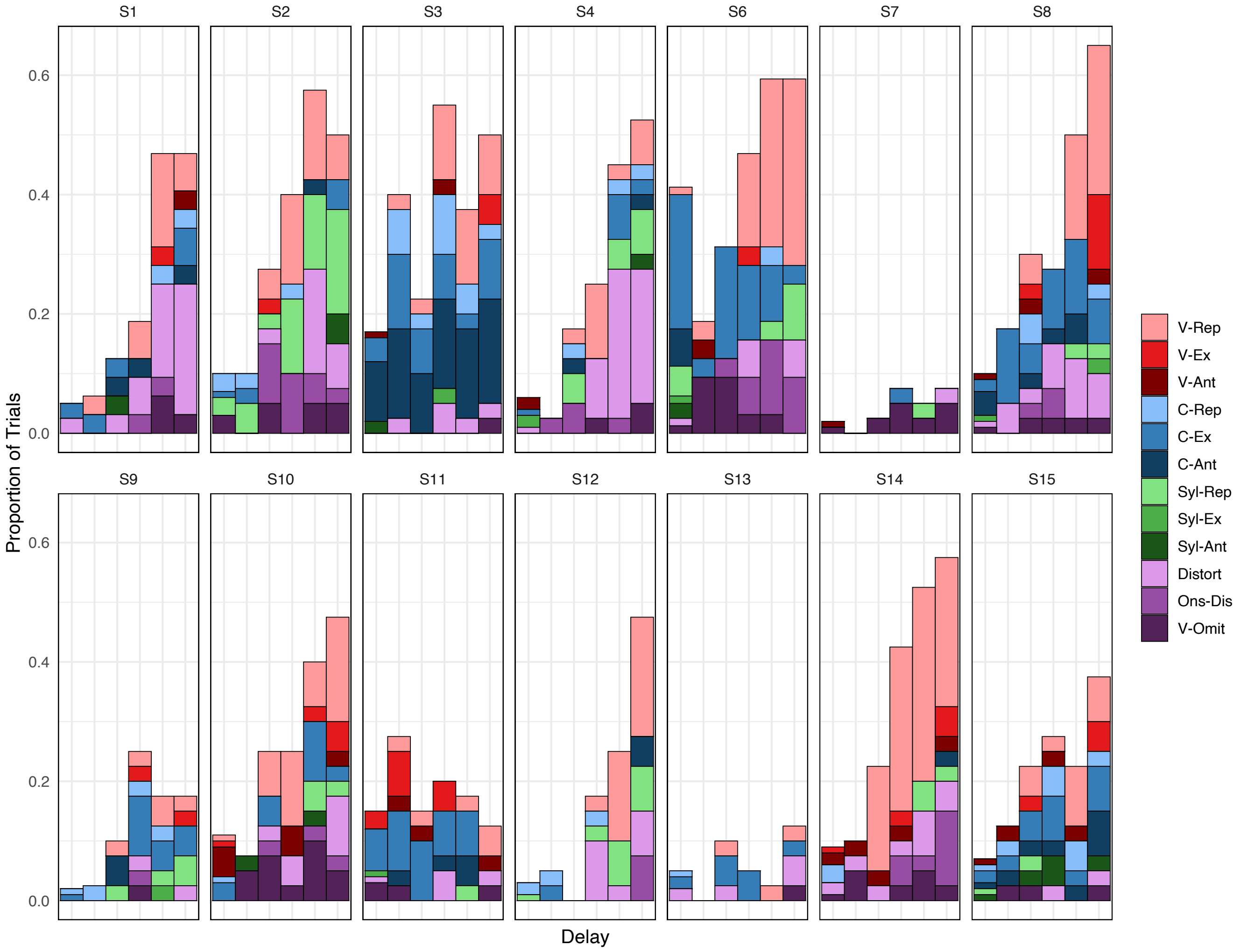

Figure 3. Speech errors produced by each participant as a function of feedback delay. Each subplot depicts the proportion of trials in which the first (if any) error produced by an individual subject was of a particular type (legend at right). The six sets of stacked bars along the x-axis correspond to the six delay latencies in increasing order: 0, 50, 100, 150, 200, 250 ms. Colors provide groupings by the sound unit involved, with red shades corresponding to vowel substitution errors, blue to consonant substitution errors, green to syllable substitution errors, and purple to errors less clearly involving discrete sound substitutions. See Table 1 for explanation of error type abbreviations.

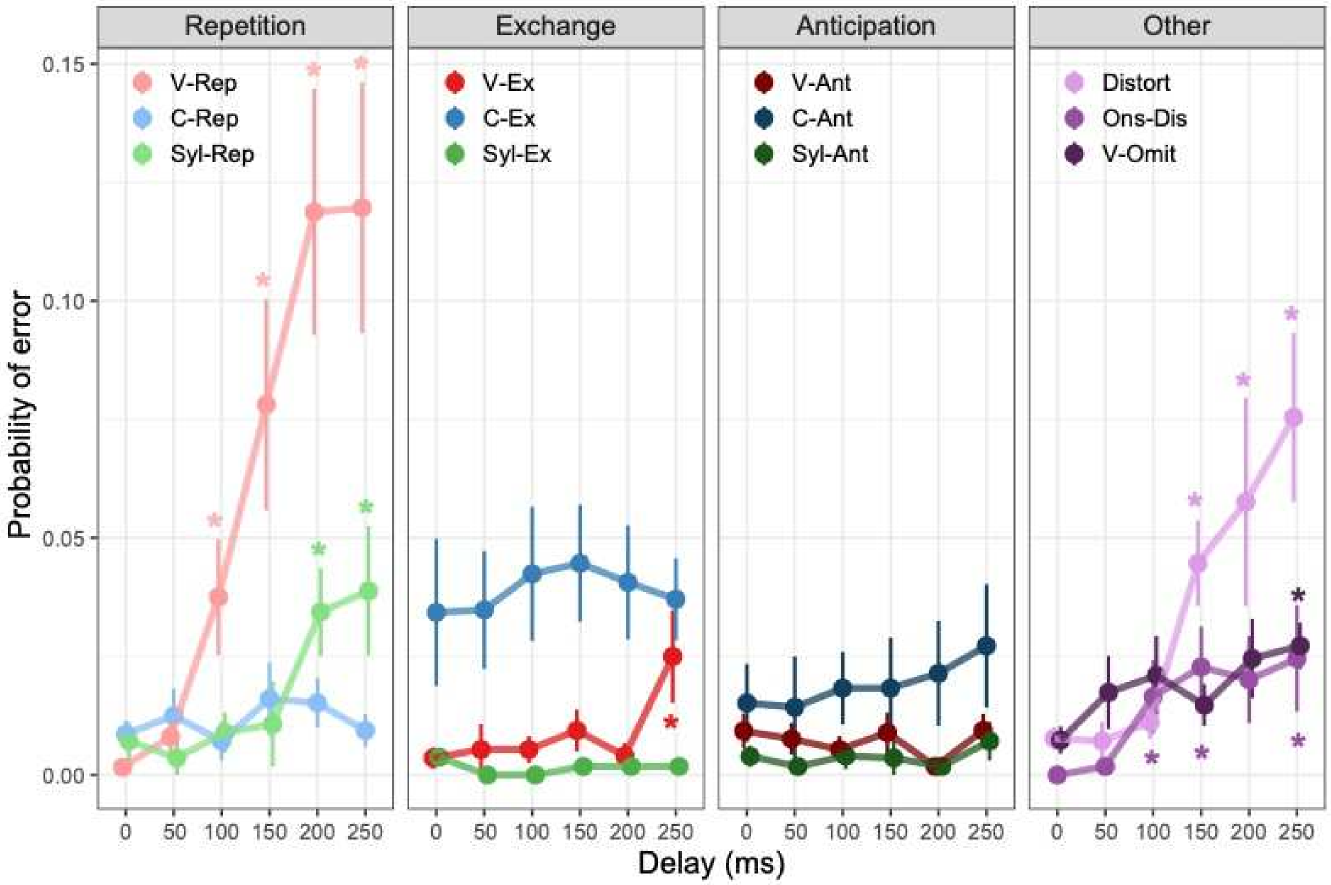

Figure 4. Mean first error rates as a function of delay. For comparison, error types are grouped into four broad categories (subplots from left to right): sound repetition errors, sound exchange errors, sound anticipation errors, and other types of errors. Lines represent the mean (across participants) fraction of trials in which the first (if any) error was of a specific type for each delay (x-axis). Error bars represent standard errors of the mean. Asterisks denote error / delay pairs where there was a significant effect relative to the zero-delay condition. Significant effects occurred for vowel repetitions (V-Rep), syllable repetitions (Syl-Rep), vowel exchanges (V-Ex), distortions, onset disfluencies (Ons-Dis), and vowel omissions (V-Omit). Note for the onset disfluency error type, the baseline level combined zero and 50 ms delay conditions to allow model convergence).

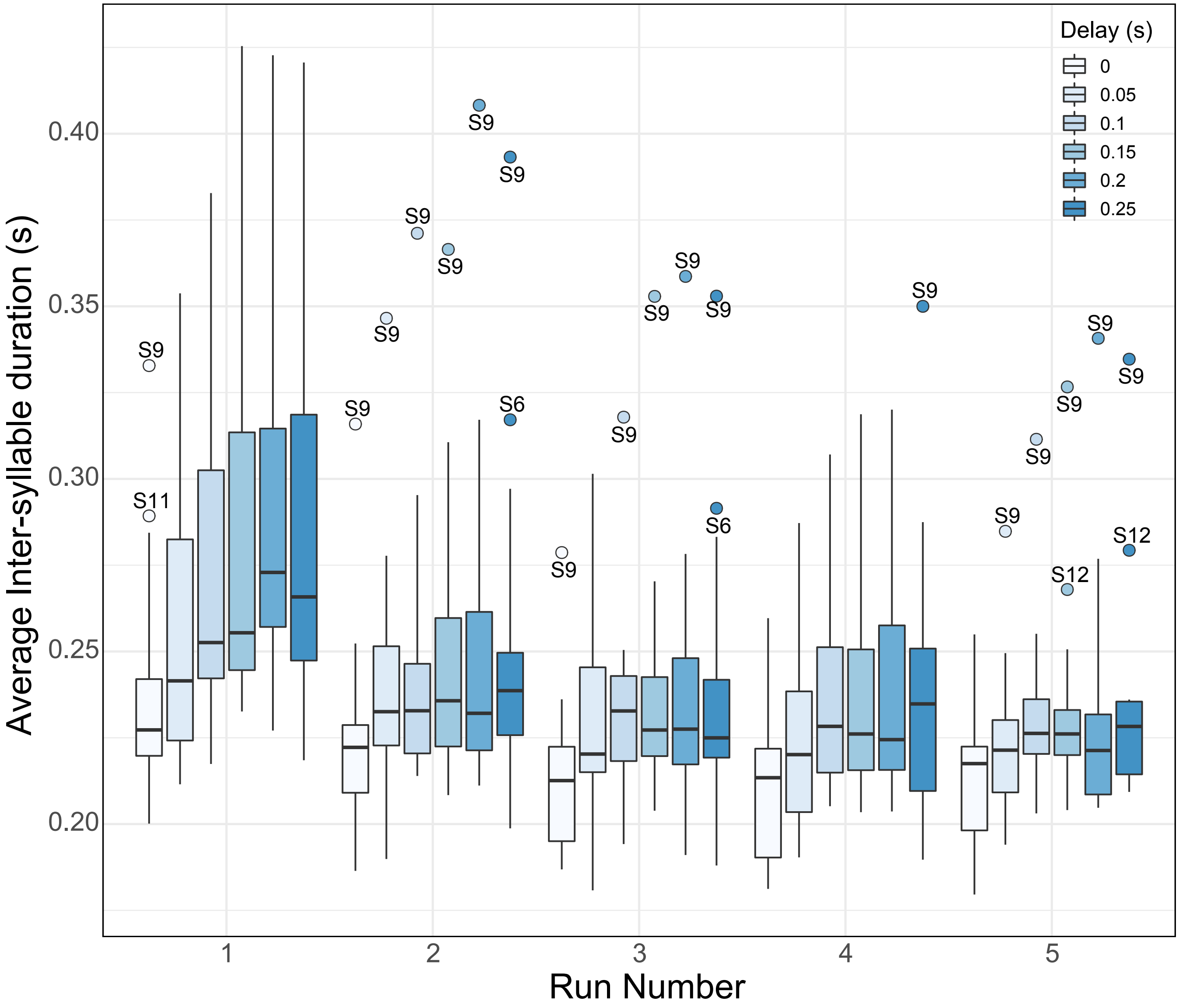

Figure 5. Mean inter-syllable durations produced as a function of run / block number and delay. Box plots show average inter-syllable onset durations, aggregated at the participant level (i.e., distributions of per-participant mean values for each run and delay). Whiskers extend up to twice the interquartile range. Outliers are labeled by participant (see e.g., Figure 2 for comparison with participant-level error data). There were significant effects of delay and run number on average inter-syllable durations, as well as a significant interaction.

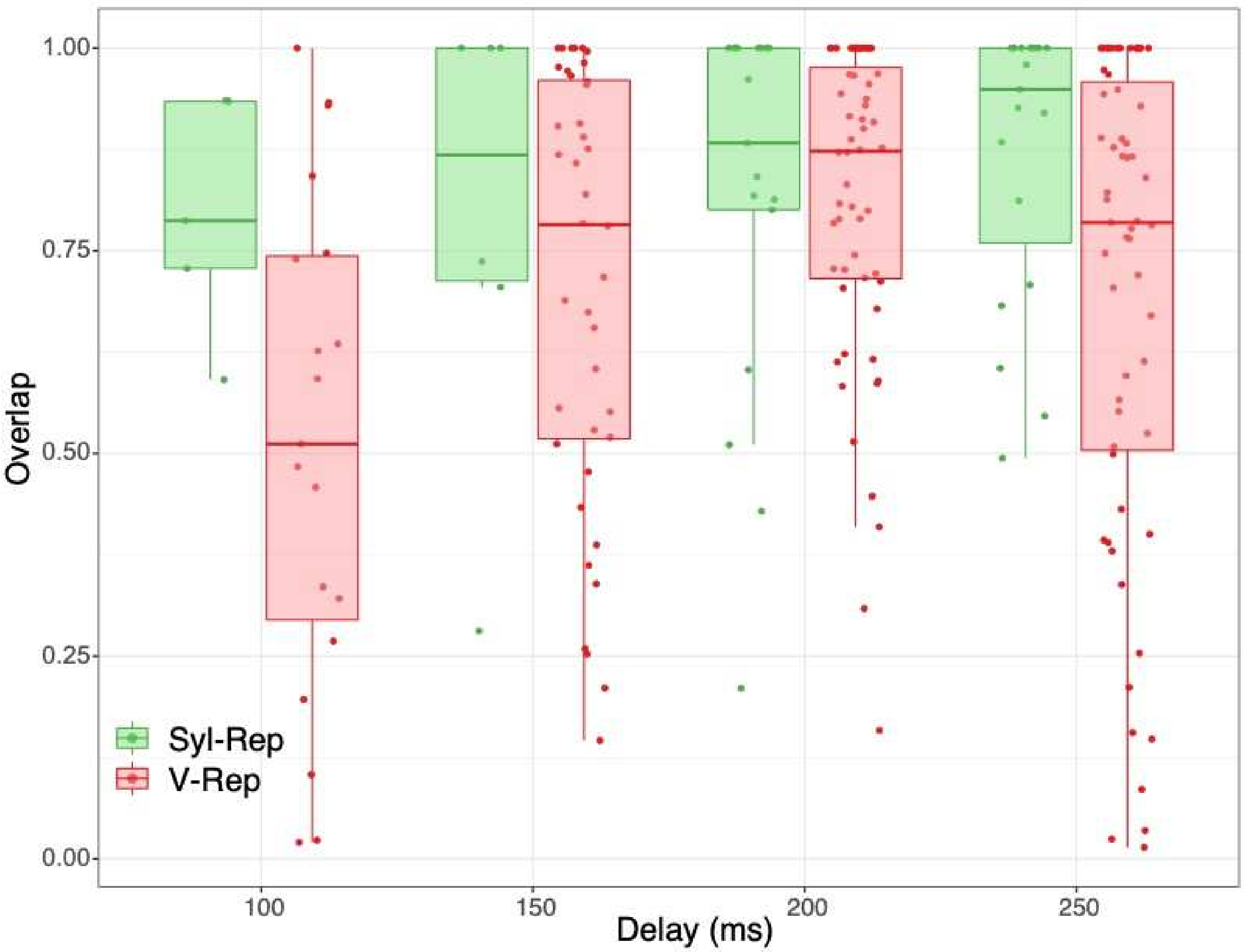

Figure 6. Temporal *overlap* measure for repetition errors. Overlap between the production time of the syllable containing the repeated unit and the time during which feedback of the previous syllable was heard. *Overlap* calculated for syllable repetition errors (Syl-Rep) is shown in green, and for vowel repetition errors (V-Rep) in red. Individual values for all errors (made by any of the participants for any delay of 100 ms or more) are indicated by dots. Note that these types of errors occurred in different proportions across delays and participants; a total of 47 syllable repetition errors and 174 vowel repetition errors were included for analysis. Overlap was significantly greater for *Syl-Rep* errors than for *Vow-Rep* errors.

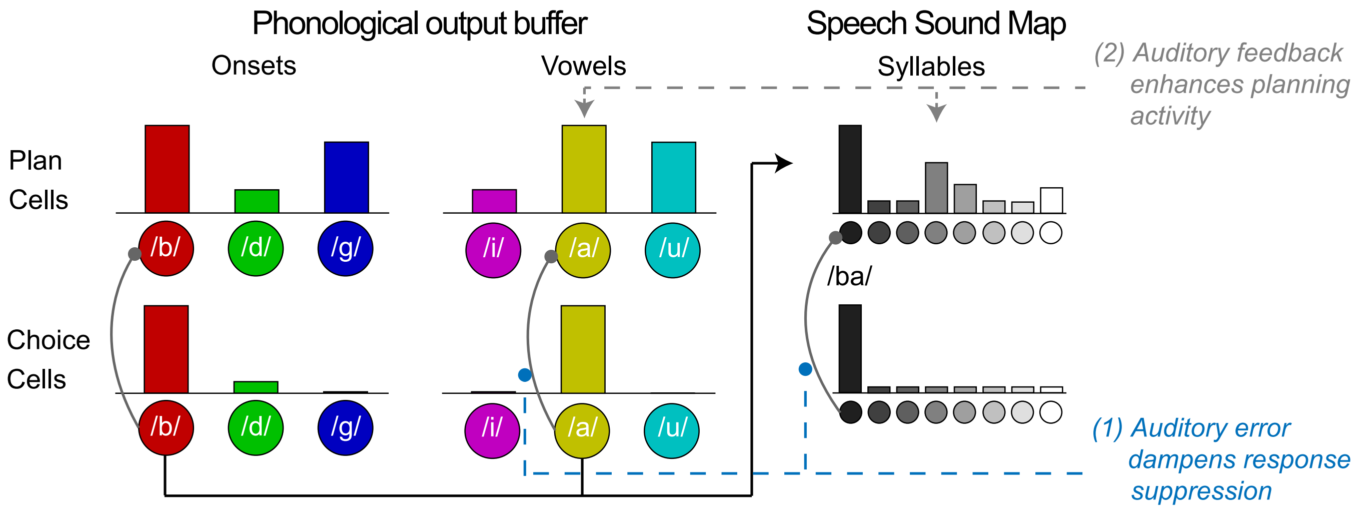
Phonological output buffer
Onsets
Vowels
Speech Sound Map
Syllables
(2) Auditory feedback enhances planning activity
Plan Cells
Choice Cells
/b/
/d/
/g/
/i/
/a/
/u/
/b/
/d/
/g/
/i/
/a/
/u/
/ba/
(1) Auditory error dampens response suppression

Figure 7. GODIVA-based schematic of possible mechanisms involved in repetition errors under DAF. The sound sequence ba-gu-di is represented by the relative activity levels in plan cells across two competitive queuing representations (one for onset consonants, one for vowels). As individual sounds are selected in the corresponding choice layers (bottom row), they suppress their planning representation (grey curved arrows). Selected sounds activate syllable plan cells in the Speech Sound Map (right, top), and a best-matching winner (/ba/) is selected for production. Dashed lines with arrows indicate the location of model inputs that could drive repetition errors either (1) through dampened response suppression signals, or (2) through enhanced activity in the planning layer due to unexpected auditory inputs. Either mechanism could operate at the level of vowels or syllables.

8 Tables

Table 1. Classification of error types used in this study.

Error type	**Sound unit**	**Code**	**Example** *Target sequence: ba-di-gu*
Repetition: a produced syllable contains an element from the syllable preceding it	Vowel	V-Rep	ba-d**a**-gu
	Consonant	C-Rep	ba-**b**i-gu
	Syllable	Syl-Rep	ba-**ba**-gu
Exchange: an element is exchanged between two adjacent syllables	Vowel	V-Ex	ba-d**u**-g**i**
	Consonant	C-Ex	ba-**g**i-**d**u
	Syllable	Syl-Ex	ba-**gu**-**di**
Anticipation: a produced syllable contains an element from the syllable that follows it in the target utterance	Vowel	V-Ant	ba-d**u**-gu
	Consonant	C-Ant	ba-**g**i-gu
	Syllable	Syl-Ant	ba-**gu**-gu
Omission: a produced syllable does not clearly contain one of the intended sound units	Vowel	V-Omit	ba-**d**-gu
	Consonant	C-Omit	ba-**i**-gu
Onset disfluency: the same onset consonant is produced 2 or more times without a clear intervening vowel	Consonant	Ons-Dis	ba-di-**g**-**g**u
Distortion: a syllable contains a phoneme that is severely distorted or not identifiable as one of the target sounds	Any	Distort	ba-dæ-gu

Table 2. Summary of generalized linear-mixed effects logistic regression model for occurrence of any speech error.

Occurrence of any error					
Random Effects	**Variance**			**Model Fit**	**Value**
Subject (Intercept)	0.66			AIC	3606.52
Sequence (Intercept)	0.07			BIC	3657.03
Fixed Effects	**Odds Ratio**	**CI (95%)**	**z**	**p**	
Intercept	0.09	0.05, 0.14	-10.02	**<0.001**	
Delay (50 ms)	1.2	0.87, 1.66	1.09	0.274	
Delay (100 ms)	2.01	1.50, 2.69	4.67	**<0.001**	
Delay (150 ms)	3.69	2.81, 4.83	9.48	**<0.001**	
Delay (200 ms)	5.17	3.97, 6.73	12.22	**<0.001**	
Delay (250 ms)	7.18	5.53, 9.31	14.83	**<0.001**	
Pairwise Contrasts	**vs. 0 ms**	**vs. 50 ms**	**vs. 100 ms**	**vs. 150 ms**	**vs. 200 ms**
Delay (50 ms)	1.20 (0.8843)				
Delay (100 ms)	**2.01 (<0.001)**	**1.67 (0.0485)**			
Delay (150 ms)	**3.69 (<0.001)**	**3.08 (<0.001)**	**1.84 (0.0012)**		
Delay (200 ms)	**5.17 (<0.001)**	**4.31 (<0.001)**	**2.58 (<0.001)**	1.40 (0.1511)	
Delay (250 ms)	**7.18 (<0.001)**	**5.99 (<0.001)**	**3.58 (<0.001)**	**1.95 (<0.001)**	1.39 (0.1423)

Table 3. Summary of generalized linear mixed effects Poisson regression model for error arrival time.

Error Arrival Time					
Random Effects	**Variance**			**Model Fit**	**Value**
Subject (Intercept)	0.041			AIC	3606.52
Sequence (Intercept)	0.013			BIC	3657.03
Fixed Effects	**Estimate**	**CI (95%)**	**z**	**p**	
Intercept	2.498	2.371, 2.625	38.569	**<0.001**	
Delay (50 ms)	-0.016	-0.103, 0.072	-0.346	0.73	
Delay (100 ms)	-0.253	-0.335, -0.172	-6.088	**<0.001**	
Delay (150 ms)	-0.352	-0.427, -0.278	-9.263	**<0.001**	
Delay (200 ms)	-0.333	-0.404, -0.261	-9.111	**<0.001**	
Delay (250 ms)	-0.409	-0.479, -0.340	-11.585	**<0.001**	
Pairwise Contrasts	**vs. 0 ms**	**vs. 50 ms**	**vs. 100 ms**	**vs. 150 ms**	**vs. 200 ms**
Delay (50 ms)	0.703 (0.9817)				
Delay (100 ms)	**5.962 (<0.001)**	**4.352 (<0.001)**			
Delay (150 ms)	**9.302 (<0.001)**	**6.941 (<0.001)**	2.461 (0.136)		
Delay (200 ms)	**9.269 (<0.001)**	**6.710 (<0.001)**	2.087 (0.294)	-0.536 (0.995)	
Delay (250 ms)	**11.769 (<0.001)**	**8.614 (<0.001)**	**4.089 (<0.001)**	1.602 (0.597)	2.309 (0.190)

Table 4. Analysis of variance (ANOVA) table using linear-mixed effects regression model for inter-syllable duration with factors *run*, *delay*, and *err*. Detailed model results, including pairwise contrasts for *run* are available in the supplemental materials.

Source	***df***	**MS**	***F***	***p***	η_p^2
run	4	0.206	365.254	< 0.001 ***	0.284
delay	5	0.117	207.661	< 0.001 ***	0.220
err (present / absent)	1	0.000	0.379	0.538	0.000
run x delay	20	0.003	5.679	< 0.001 ***	0.030
residuals	3740	0.001			

Table 5. Linear mixed effects model estimates for the *overlap* measure examined during syllable and vowel repetition errors.

Overlap timing measure					
Random Effects	**Variance**			**Model Fit**	**Value**
Subject (Intercept)	0.076			AIC	865.888
Fixed Effects	**Coefficient**	**CI (95%)**	**t-value**	**dof**	**p**
Intercept	1.490	1.136, 1.843	8.26	19.147	**<0.001**
Delay (150 ms)	-0.683	-1.328, -0.038	-2.077	212.723	**0.039**
Delay (200 ms)	0.054	-0.529, 0.637	0.182	212.793	0.856
Delay (250 ms)	0.395	-0.040, 0.830	1.781	210.785	0.076
Error unit type (syl vs. vow rep)	0.406	0.091, 0.721	2.529	181.910	**0.012**
Delay (150 ms) x Error unit type	0.296	-0.347, 0.939	0.901	212.844	0.368
Delay (200 ms) x Error unit type	-0.145	-0.727, 0.437	-0.488	213.000	0.626
Delay (250 ms) x Error unit type	-0.299	-0.734, 0.137	-1.344	211.484	0.180

9 Supplemental Files

Supplemental File 1: This Excel Spreadsheet provides multiple worksheets, each of which provides a tabular summary for an individual generalized linear mixed-effects logistic regression model for one error type (see Section 3.2 Error patterns by type). Tables describe fixed and random effects and provide pairwise contrasts for delay.

Supplemental File 2: Summary of coefficients for linear mixed effects model estimating the effects of delay, run number, and presence/absence of an error on inter-syllable durations (see also main Table 4).

Supplemental File 3: Summary of coefficients for linear mixed effects model estimating temporal measures *onset interval* and *peak interval* during syllable and vowel repetition errors (see also main Table 5 for *overlap* measure).

Supplemental Figure 1: Sound substitutions errors produced by each participant as a function of feedback delay. This figure shows a subset of the errors shown in Figure 3.

Supplemental Figure 2: Other (not sound substitution) errors produced by each participant as a function of feedback delay. This figure shows a subset of the errors shown in Figure 3.

Supplemental Figure 3: Interaction plot showing model estimated means (and 95% confidence intervals) for inter-syllable durations as a function of delay and run number.