JASA ARTICLE

Compensation to real-time temporal auditory feedback perturbation depends on syllable position

Miriam Oschkinat[a] and Philip Hoole
Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Schellingstrasse 3, Munich, 80799, Germany

ABSTRACT:
Auditory feedback perturbations involving spectral shifts indicated a crucial contribution of auditory feedback to planning and execution of speech. However, much less is known about the contribution of auditory feedback with respect to temporal properties of speech. The current study aimed at providing insight into the representation of temporal properties of speech and the relevance of auditory feedback for speech timing. Real-time auditory feedback perturbations were applied in the temporal domain, viz., stretching and compressing of consonant-consonant-vowel (CCV) durations in onset + nucleus vs vowel-consonant-consonant (VCC) durations in nucleus + coda. Since CCV forms a gesturally more cohesive and stable structure than VCC, greater articulatory adjustments to nucleus + coda (VCC) perturbation were expected. The results show that speakers compensate for focal temporal feedback alterations. Responses to VCC perturbation were greater than to CCV perturbation, suggesting less deformability of onsets when confronted with temporally perturbed auditory feedback. Further, responses to CCV perturbation rather reflected within-trial reactive compensation, whereas VCC compensation was more pronounced and indicative of adaptive behavior. Accordingly, planning and execution of temporal properties of speech are indeed guided by auditory feedback, but the precise nature of the reaction to perturbations is linked to the structural position in the syllable and the associated feedforward timing strategies. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001765

(Received 14 February 2020; revised 27 July 2020; accepted 29 July 2020; published online 15 September 2020)
[Editor: Susanne Fuchs] Pages: 1478–1495

I. INTRODUCTION

Human speech is a unique auditory-motor communication mode that involves a wide set of physiological, neurological, and behavioral contributors. In research on planning, production, and perception of speech, the connection and interaction of these contributors have been of key interest.

As part of this, perturbations of auditory feedback have proven very useful for studying the contribution of self-perception to planning and control of speech. A diverse body of research has shown that subjects adjust productions within a short time frame when the auditory feedback of their own speech is altered. In the manipulation of the fundamental frequency (Jones and Munhall, 2000; Xu *et al.*, 2004; Patel *et al.*, 2011), formant frequencies of vowels (Houde and Jordan, 1998, 2002; Purcell and Munhall, 2006a,b; Villacorta *et al.*, 2007; MacDonald *et al.*, 2010; MacDonald *et al.*, 2011; Mitsuya *et al.*, 2011), or center of gravity (CoG) of fricatives (Shiller *et al.*, 2009; Casserly, 2011; Klein *et al.*, 2019), responses were mainly exhibited in the opposite direction to the applied shift, causing *compensation* for the received feedback. While spectral alterations have been extensively studied, much less is known about the impact of focal temporal auditory feedback alterations on speech production. The current study aims at filling this gap by applying auditory feedback perturbations in the temporal domain with a specific focus on different prosodic positions within the syllable.

Spectral auditory feedback perturbations revealed reactions on different levels in response to applied shifts. While some studies found compensatory responses in the control of ongoing speech movements (*online compensation*), others investigated the effects of compensatory *adaptation* for perturbed segments. Adaptation is a (compensatory) response that indicates a modification of the underlying representations at the planning level of motor control, mostly notable in a persistence of articulatory adjustments when normal feedback is restored or a transfer of articulatory adjustments to other (not perturbed) sounds of similar quality or in a similar context (Houde and Jordan, 1998, 2002; Villacorta *et al.*, 2007; Caudrelier *et al.*, 2016). Online compensation and adaptation have mainly been elicited in two different experimental paradigms. While some studies applied unexpected feedback shifts in a small number of random trials to interfere with the online control of speech, others used consistently perturbed feedback, thus targeting the predictions about properties of speech sounds. With unexpected, randomly applied perturbations, compensatory responses were found with a latency typically between ∼120 and 200 ms after perturbation onset (Burnett *et al.*, 1998; Donath *et al.*, 2002; Xu *et al.*, 2004; Purcell and Munhall, 2006b; Tourville *et al.*, 2008; Niziolek and Guenther, 2013). This reaction indicates that the motor system is capable of adjusting online in moment-to-moment control during the execution of sustained vowels or more complex sound patterns,

[a]Electronic mail: Miriam.Oschkinat@phonetik.uni-muenchen.de

 © 2020 Acoustical Society of America

https://doi.org/10.1121/10.0001765

such as syllables, but with a delay caused by the latency of sensory feedback in feedback-feedforward loops. With the other paradigm of consistent perturbation and compensatory reactions after a period of training, a continuous mismatch between predictions and actual received feedback leads to a modification of the underlying motor plan. The latter method can trigger more local compensatory responses that take effect exactly at that part of the speech signal that has been perturbed. Thus, predictions are made (or updated) based on previous trials, bypassing the fact that auditory feedback is too slow for closed-loop online control (Purcell and Munhall, 2006a).

Together, the two compensatory mechanisms give insight into the involvement of auditory feedback at different levels of speech production. While online compensation indicates a link between auditory feedback and the control of ongoing speech, adaptation speaks for an involvement of auditory feedback in the establishment and tuning of feedforward mechanisms. To date, several approaches to modeling speech production that incorporate a link between auditory feedback and the control level can account for online compensation to altered auditory feedback, like the *DIVA* model (Guenther *et al.*, 2006; Tourville and Guenther, 2011), *state feedback control* (Houde and Nagarajan, 2011; Houde *et al.*, 2014; Houde and Chang, 2015), or the *FACTS* model (Ramanarayanan *et al.*, 2016; Parrell *et al.*, 2018; Parrell *et al.*, 2019b). The explanation of adaptation effects, however, demands an integration of auditory feedback into mechanisms at the planning level, as incorporated in the DIVA model, the *ACT* (*ion-based model of speech production*; Kröger *et al.*, 2009), or more recent versions of *GEPETTO* (Patri *et al.*, 2018; Patri *et al.*, 2019). One of the most comprehensive approaches to modeling speech production, and one that is able to account for both online compensation and adaptation, is the DIVA model.

DIVA hypothesizes spatio-temporal target regions for phonemes or syllables spanning auditory and somatosensory dimensions. The sensory feedback serves to monitor and evaluate the quality of the produced sound. If a production is, for example, spectrally not located within the auditory target dimensions of the desired speech sound, the commands for articulatory movements in the current or following productions will be updated to better match the desired target. If the mismatch persists, the target dimensions can be adjusted eventually. While the results of *spectral* auditory feedback perturbation constitute strong support for the DIVA framework, there is not much evidence about how *temporal* properties of speech, such as duration of sounds and the relation between them within syllables, are established and controlled. In many approaches to modeling speech production, temporal properties of speech are either modeled as fixed but include auditory feedback (as in DIVA, recent versions of GEPETTO, or ACT), or the control of speech timing is modeled dynamically but exclusively through feedforward mechanisms, as in the articulatory phonology/task-dynamics framework or the FACTS model (see Parrell *et al.*, 2019a, for an overview of current models of speech motor control). It is true that task dynamics assumes the availability of somatosensory feedback for error correction at the interarticulator level. However, this feedback-based correction operates in task-space with no feedback connection to the intergestural level, where context-independent timing relations and gestural activation patterns are represented.[1]

The coupling of action and perception specifically for timing mechanisms has been investigated comparatively infrequently in speech sciences but has experienced a broad focus of interest in cognitive sciences and music research. The anticipation and precise timing of motor execution, termed *predictive timing* (Debrabant *et al.*, 2012), has mainly been studied through, for example, the coordination of rhythmic motor action to an external beat (Repp and Su, 2013, for an overview). In such tasks, an internal prediction of timing is generated and updated with increasing success in matching the auditorily received beat. Turning back to speech production, it seems that also here, planning and execution comprise predictions about the time and time frame of a particular speech sound (Kotz and Schwartze, 2010). Further evidence for this assumption is provided by research on people who stutter: While people who stutter show an impairment in the precise timing of speech sounds, particularly in syllable onsets (see, e.g., Hubbard, 1998; Max and Gracco, 2005; Etchell *et al.*, 2014), they also show deficits in nonspeech predictive timing tasks, such as tapping to a beat (Falk *et al.*, 2015).

For a better understanding of predictive timing mechanisms in speech, focal temporal auditory feedback perturbation should give insight into the monitoring of speech timing and the flexibility of the motor system to update temporal representations. Cai *et al.* (2011) examined the online control of speech timing by disrupting the temporal fine structure of an utterance with temporally altered auditory feedback. They altered the F2 minimum of the vowel [u] in "owe" within the utterance "I owe you a yo-yo." In one perturbation condition, the F2 minimum was either accelerated, whereby it was perceived earlier in time, while in another condition it was decelerated, eliciting a later percept of the vowel target. They found reactions in the same direction as the perturbation for the deceleration condition (global delaying/lengthening of following segments). However, there is no clear indication of what a specific adjustment in the other direction would comprise: Keeping in mind the general reaction latency to unpredicted perturbations, an anticipation of the following segments as a reaction to the unexpected temporal perturbation might have been rather improbable in our opinion. Certainly, Cai *et al.* (2011) were able to show that subjects react to an unpredicted perturbation of perceived timing. With the global delay in reaction, however, their study could not directly give information about temporal representations of specific speech sounds nor indicate a specific compensatory behavior.

Taking this into account, we make the general assumption that online compensation to focal temporal perturbation is not possible. Unlike spectral properties of speech that

evolve over time, temporal dimensions (e.g., sound durations) cannot be adjusted instantly within the ongoing production since the duration of a segment is not determinable until it has been perceived in its entirety.

A different approach to altering speech timing is found in the study by Mitsuya *et al.* (2014). Their study altered contrastive phonation timing of voice onset time (VOT) with an adaptation paradigm of persistent and constant perturbation. Subjects either produced the word "dipper" or the word "tipper" while receiving a prerecorded version of their own productions of the other token. Unlike Cai *et al.* (2011), the total duration of a sound segment (VOT of the initial plosive) was altered in the auditory feedback, although not in real-time. They found adaptive compensation of around 15%–20% for the perturbed segments, indicating that auditory feedback plays a role in temporal planning of phonation. However, as subjects were receiving prerecorded tokens, their compensation did not actually have any effect on the perceived outcome. Very recently, the study by Floegel *et al.* (2020) combined both spectral and temporal real-time auditory feedback perturbations with functional magnetic resonance imaging (fMRI). With a real-time adaptation paradigm, they stretched single sounds in monosyllabic words whereby subjects compensated with a shortening of the perturbed segments.

In previous spectral or temporal perturbations, while vowels and consonants at different locations within the syllable have been perturbed, prosodic functions of the different parts of the syllable have nonetheless not been taken directly into consideration as an influencing factor. In temporal perturbation of fluent speech, there are, however, good reasons to assume that prosodic functions of different parts of the syllable could be highly influential for the behavioral reaction. Notably, the articulatory phonology/task dynamics framework has elaborated different timing and coordinative patterns for segments as a function of the syllable position.

With respect to the syllable structure, intergestural timing was modeled with *coupled planning oscillators*, which may couple mainly in-phase or antiphase with each other in fluent speech (Goldstein *et al.*, 2009; Nam *et al.*, 2009). Hereby, different coordinative relations (coupling topologies) between gestures were found for onset vs coda position. Onsets are coupled antiphase with each other but in-phase with the vowel to form a global coordination structure, while vowel + coda segments constitute rather local patterns of coordination, each being coupled antiphase with the preceding sound. The in-phase coupling with the vowel exhibited in onsets is assumed to represent a more stable coupling topology than the purely local coupling in the coda, which allows for higher variability in timing of codas but constitutes greater articulatory stability for onsets (Byrd, 1996; Browman and Goldstein, 2000; Goldstein and Pouplier, 2014).

The current study aims at testing how coupling concepts of speech timing anchored in feedforward mechanisms might combine with the idea that auditory feedback interacts with the planning and control of speech timing. More specifically, using a temporal auditory feedback adaptation paradigm, absolute durations of sounds with different functionality for syllable timing will be stretched and compressed in real-time.

Based on this consideration, we are led to a design with two experimental conditions: First, manipulations are applied to onset + vowel (CCV) in consonant-consonant-vowel-consonant (CCVC) syllable (onset condition), and second to vowel + coda (VCC) in a CVCC syllable with similar phonological and lexical context (coda condition). We predict durational adjustments of the perturbed segments in the opposite direction to the applied shift. Since onset + vowel sequences show greater temporal stability in feedforward control than vowel + coda, we expect them to be less malleable in the face of an auditory perturbation.

The manipulation we present in this study can thus be expected to give further insight into potentially different underlying timing mechanisms related to different structural locations in the syllable. We believe that the influence of such structural considerations on the malleability of motor representations is a neglected issue in perturbation studies in general, and, as we have argued above, is likely to be particularly relevant specifically in the field of temporal perturbations. By employing consistent perturbations that can be expected to become predictable for the subject, we can study compensatory reactions exactly at the perturbation location itself and consequently shed light on the representation of temporal properties of the individual speech sound. In addition to the focus on syllable structure, further motivation for the present study is given quite simply by how little is known about the extent to which temporal properties of speech follow similar mechanisms in speech planning to those for spectral/spatial properties.

The studies of Cai *et al.* (2011), Mitsuya *et al.* (2014), and Floegel *et al.* (2020) all lead to the general expectation that subjects are indeed sensitive to focal temporal auditory feedback perturbation and the studies of Mitsuya *et al.* (2014) and Floegel *et al.* (2020) (again, in analogy to spectral perturbations) lead to the general expectation that subjects show compensatory durational adjustments. However, these two studies (of particular relevance to our own) were quite naturally only able to address compensatory behavior in a small subset of potentially relevant contexts: Mitsuya *et al.* (2014) looked at a specific subsegmental phonological contrast in single disyllabic words, and Floegel *et al.* (2020) stretched single sounds in isolated monosyllables. Thus, essentially nothing is known about how additional possible prosodic and segmental contexts may affect compensatory behavior. Our study aims to contribute to this more general understanding by investigating the effect of a more complex bidirectional perturbation applied to multiple segments within a syllable, which, in turn, is part of a complete multisyllabic phrase.

II. METHODS

A. Speech material and subjects

The experimental setup was geared to enable real-time auditory feedback alterations to a CCV sequence (onset condition) and a VCC sequence (coda condition), both with

https://doi.org/10.1121/10.0001765

similar phonological context and lexical frequency. Therefore, for the onset perturbation condition, the German word "Pfannkuchen" (/ˈpfanku:xən/, *pancake/s*) was chosen, and for the coda perturbation condition, the German word "Napfkuchen" (/ˈnapfku:xən/, *ring cake/s*) was chosen. The first syllable of each word ("Pfann" /pfan/ or "Napf" /napf/, respectively) was the focus of interest for manipulation. Manipulations covered the onset consonants and the vowel (/pfa/) in the onset condition and the vowel and the coda consonants (/apf/) in the coda condition. Unlike spectral perturbations, where a defined amount of upward or downward spectral shift can be systematically applied to the signal, the creation of real-time temporally altered feedback of multisyllabic speech holds the constraint that it is mandatory to first stretch segments before compressing others. With only a stretching of segments, the following signal would be perceived as overall delayed, whereas compression on its own is not possible because, in this case, the signal needed as feedback would not yet have been produced.

For the present experiment, the component durations of the CCV and VCC sequences (/pfa/ for the onset condition and /apf/ for the coda condition) were, respectively, stretched (first 50% of the sequence) and compressed (second 50% of the sequence) and fed back almost in real-time. Hence, in the onset condition the onset consonants (CC (/pf/) were mostly stretched and the vowel (/a/) compressed, whereas in the coda condition, the vowel (/a/) was stretched and the coda consonants (CC /pf/) were mostly compressed. The amount of perturbation was in proportion to the individually produced segment length and, hence, not equal in absolute duration over all subjects. Examples of perturbation for both onset and coda conditions can be found in Figs. 1(A) and 1(B), respectively.

The test words were spoken after the carrier word "besser" (/ˈbɛsɐ/, *better*), resulting in the German phrases "besser Pfannkuchen" or "besser Napfkuchen."

Forty-five monolingual native speakers of German between 19 and 30 years of age (mean age, 23 years old, 34 females) participated in both experiment conditions, the onset and the coda manipulation. The order of testing was counterbalanced over subjects. None of them claimed to have any speech or voice disorder nor any hearing impairments. Subjects were compensated for their participation.

B. Experimental setup

The experiment was conducted in MATLAB (The MathWorks Inc., Natick, MA) using the AUDAPTER software package of Cai *et al*. (2008). Originally developed for formant manipulations in utterances with continuous voicing, further versions allow for delay shifts, time warping, and pitch shifts in all kinds of utterances (Cai *et al*., 2011; Tourville *et al*., 2013). Audapter is coded in C++ and implemented in MATLAB for configurable real-time manipulation of acoustic parameters of speech. The software package includes both the core algorithms for real-time speech signal processing and, additionally, wrap-arounds in MATLAB supporting psychophysical experiments (Cai, 2014).

Because the perturbation is supposed to target a preselected part of an utterance, there is a need for an online status tracking (OST), which contains a set of heuristic rules to recognize specific segments in speech. The OST is based on detection of user-configurable predefined high- and low-frequency weighted intensity thresholds based on the speech signal's short-time root-mean-square (RMS) amplitude. In this experiment, the end of the OST marks the start of the perturbation section where the manipulation is applied. OST thresholds were set up to track the single phonemes in the word "besser" (/bɛsɐ/). The onset of the /ɐ/ was the last automatically tracked OST state. From the onset of the /ɐ/ to the onset of /p/ in Pfannkuchen or the onset of /a/ in Napfkuchen, an individual amount of *elapsed time* was implemented per subject as a final individual OST state. To estimate the amount of *elapsed time* and the length of the perturbation section (the length of the CCV and VCC sequences), each subject underwent a pretest per experiment condition that comprised 15–20 productions of the desired utterance without perturbation. These trials served as practice to produce the sequence naturally and at a constant speech tempo. Subsequently, the experimenter measured the mean *elapsed time* and the mean CCV (/pfa/) or VCC (/apf/) duration from the pretest trials and embedded those into the test procedure as the final OST state and the time frame for the perturbation section. Before the testing started, one token that was the closest to the mean *elapsed time* and mean CCV/VCC measure was presented to the subject as an example token for their speaking rate.

Subjects wore E-A-RTone™ 3A in-ear earphones with E-A-RLINK foam eartips (3M, Saint Paul, MN) for perturbed feedback and a Sennheiser H74 headset microphone (Wedemark, Germany) placed 3 cm from the corner of the mouth. The foam eartips ensure that the manipulated feedback rather than the airborne sound is predominantly perceived and also minimizes the occlusion effect [see Fig. 1(C) for the setup]. Subjects spoke the target phrase ("besser Pfannkuchen" or "besser" Napfkuchen) 110 times per condition. The phrase was lexically presented on a screen, and the time span of recording was indicated by a green frame around the target phrase. The duration of each recording was set to 2.5 s, which allowed the subjects to choose an individual comfortable and natural speaking rate without providing too much time for high variability in the speaking rate within and between subjects. Throughout the experiment, subjects were required to keep their speech rate as constant as possible. The spoken signal was fed through a MOTU MicroBook II (Cambridge, MA) to the computer, where the perturbation algorithm was applied. The manipulated signal was then sent back through a PreSonus Monitor Station (Baton Rouge, LA) and amplified via a PreSonus HP4 headphone amplifier before it reached the subject's ears with a total delay of not more than 24 ms. The playback volume was set to a comfortable level but loud enough to ensure that they did not hear their own airborne sound. The

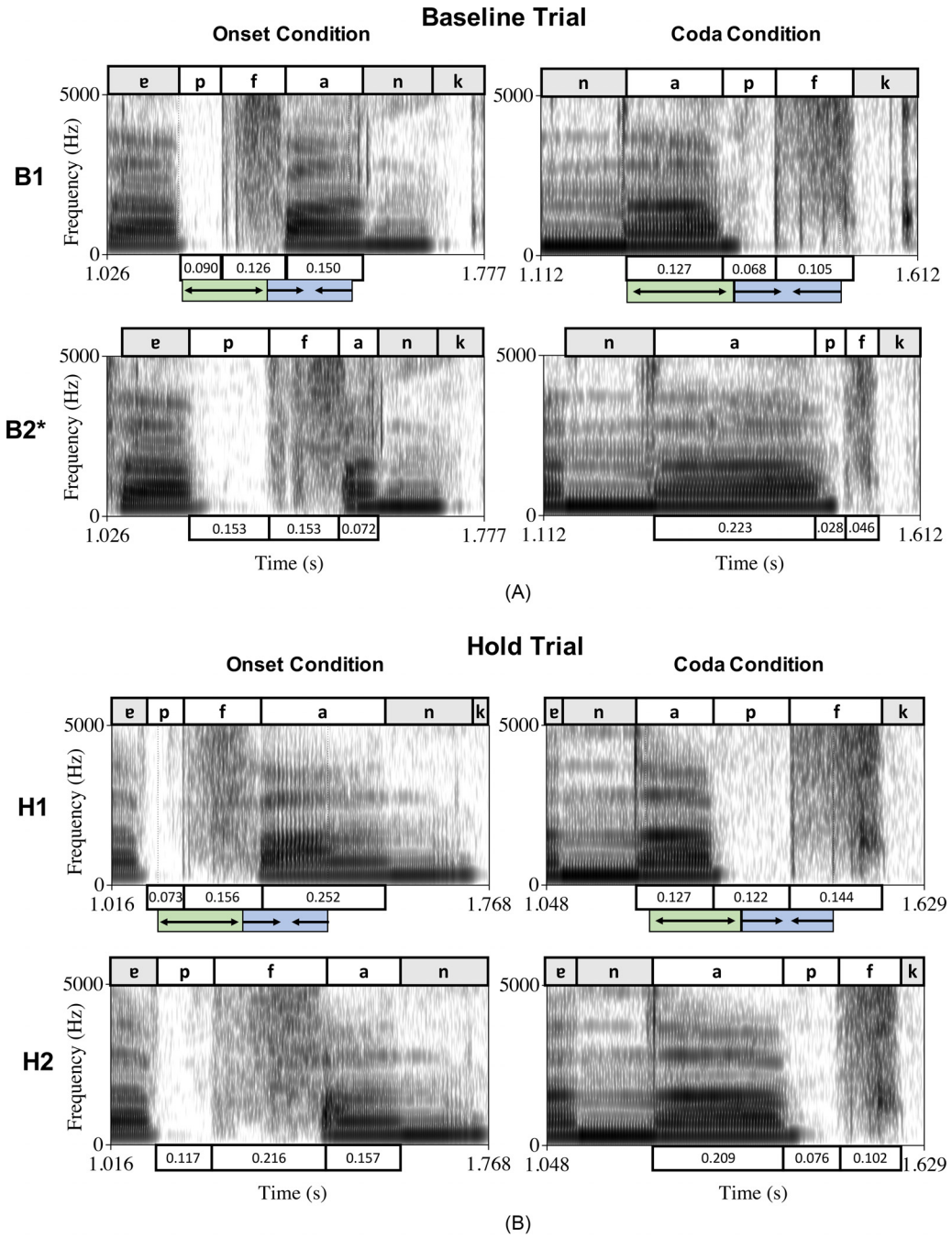

FIG. 1. (Color online) (A) Spectrograms of a baseline trial per condition of one subject. The onset perturbation condition appears in the left panels ("bess**er Pfannk**uchen," bold section visible in the spectrograms) and the coda perturbation condition appears in the right panels ("bess**er Napfk**uchen," bold section visible in the spectrograms). The upper panels show the produced signal of the baseline trial (B1), and the lower panels show a simulated maximum perturbation of the same trial (B2*). The simulation of the perturbation in the baseline visualizes the perturbation of a trial that is not already produced with articulatory adjustments to the perturbation and gives a "clean" indication of full perturbation. Segments of interest are marked above the spectrogram with their durations shown below the spectograms. The green-blue bars below the upper spectrograms mark the perturbation section. The signal comprising the first half of the perturbation section was stretched (green bar) and the signal in the second half of the perturbation section was compressed (blue bar), resulting in the sound durations in the panel below (B2*). Note that the perturbed signal includes the Audapter delay of 24 ms. (B) spectrograms of a hold trial per condition of the same subject as in (A). H1 shows the produced signal of a hold trial, and H2 shows the perturbed feedback of the same trial. The onset perturbation condition appears in the left panels (bess**er Pfannk**uchen, bold section visible in the spectrograms) and the coda perturbation condition appears in the right panels (bess**er Napfk**uchen, bold section visible in the spectrograms). Segments of interest are marked above the spectrogram and their durations are marked below the spectrogram. Note that productions in the upper panels might already be produced compensatorily. The green-blue bars below the upper spectrograms mark the perturbation section. The signal comprising the first half of the perturbation section was stretched (green bar) and the signal in the second half of the perturbation section was compressed (blue bar), resulting in the sound durations in the panel below (H2). Note that the perturbed signal includes the Audapter delay of 24 ms. (C) Experimental setup. (D) Visualization of the four phases of the experiment and the applied perturbation in each phase. The green line visualizes the stretching and the blue line visualizes the compression.

https://doi.org/10.1121/10.0001765

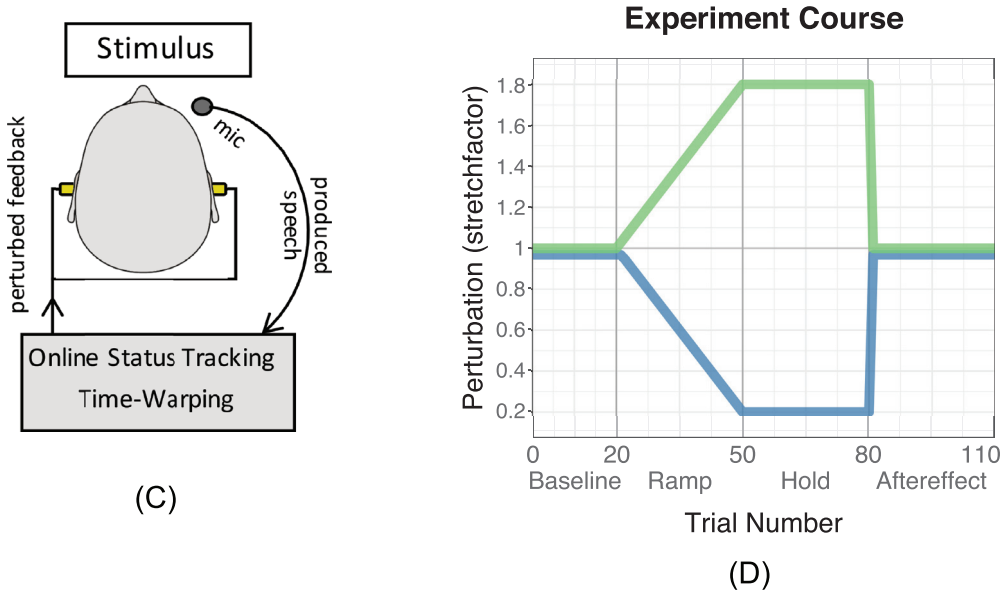

FIG. 1. (Color online) (*continued*)

level was based on tests with pilot subjects and was kept constant for all further subjects with an adequate modulation of the microphone level for each subject's speech. Subject and experimenter were able to communicate during the whole session.

Perturbation was applied in phases with different perturbation magnitudes as was done in previous studies (e.g., Purcell and Munhall, 2006a). First, there was a baseline with no perturbation (20 trials), followed by a ramp phase with gradually increasing perturbation (30 trials), after that, the maximum amount of perturbation was held for another 30 trials (hold phase), and the experiment was completed after 30 further trials with no perturbation again [aftereffect phase; Fig. 1(D)]. In the hold phase with maximum perturbation, the first half of the perturbation section was stretched to 1.8 times the input duration while the second half of the perturbation section was compressed to 0.2 times the input duration.

III. ANALYSES

A. Data handling

For the analyses, all trials with dysfluencies or slips of the tongue and utterances that exceeded the recording window were discarded ("rubbish trials"). Per subject and perturbation condition, all ramp and hold trials in which the perturbation of the vowel /a/ or the CC segment /pf/ did not take effect in the intended perturbation direction (caused, e.g., by a malfunction of tracking, a poor fit of the perturbation section, or a high variance in speaking rate) were excluded with an automated MATLAB script. Subjects with less than 16 out of 30 acceptable hold trials were excluded from following calculations; hence, the number of hold phase trials varied between 30 and 16 trials per subject. Visual examination of the data indicated that with a minimum of 16 perturbed trials, the number of available trials did not cause any systematic effects. After excluding subjects with less than 16 acceptable hold phase trials, data were available for 34 subjects for the onset condition (mean, 23 years old; 27 female) and 33 subjects for the coda condition (mean, 23 years old; 27 female). Twenty-eight of those subjects provided data for both perturbation conditions. From a total of 3740 trials in the onset condition (34 subjects × 110 trials), 166 trials were discarded (rubbish trials, 14; poor fit of the perturbation section, 152). In the coda condition from 3630 trials (33 subjects × 110 trials), 149 trials were excluded (rubbish trials, 18; poor fit of the perturbation section, 131).

The majority of female subjects is mainly caused by the discrepancy in the readiness to participate in experiments in the tested environment. To our knowledge, there is no study that provides evidence for a sex-related difference in perception of auditory feedback and integration into the speech motor plan for fluent speech (but see Chen *et al.*, 2010, for pitch in sustained vowels). Hence, the mentioned discrepancy is not expected to cause a systematic sex-related effect in this study.

B. Measures

Durations of each phonological segment of the spoken utterance were defined and measured manually in PRAAT (Boersma and Weenink, 1999). Subsequently, the measured durations were normalized by word duration ("Pfannkuchen" or "Napfkuchen"). Differences in normalized durations rather reflect changes in duration of segments within the word as opposed to changes in speaking rate (e.g., an overall slowing down or speeding up during the experiment would show differences in absolute segment durations but does not necessarily indicate a duration difference of the segment within the word). In previous studies, the first trials were often excluded due to higher variance in speaking at the beginning of the experiment

(e.g., Mitsuya *et al.*, 2014, excluded the first ten trials). In the current study, higher variability in production during the first nine trials was observed. Therefore, for all subjects the first 9 baseline trials were discarded, resulting in 11 baseline trials. A baseline mean was calculated over those trials and the normalized durations were referenced to this baseline mean, further referred to as *normalized relative durations*.

Motivated by the hypotheses of the current study, the following analyses focus on two segments per perturbation condition, the CC segment /pf/ and the vowel /a/. Since it is conceivable that the single CC consonants show individual reaction patterns, the CC segment will subsequently be broken down into its components (C1 /p/ and C2 /f/), although we have no clear hypothesis about their individual behaviors. Figure 2 visualizes the produced normalized relative durations averaged over all subjects of the CC segment /pf/ (green dots) and the vowel /a/ (blue rhombuses). The baseline mean (calculated from trials 10–20) represents the 0 line. Positive values indicate a lengthening and negative values indicate a shortening, relative to baseline productions. The spoken signal is shown in solid colors, the perturbed (heard) signal is shown with higher transparency. Please note that the perturbed/heard signal does not represent a one-to-one mapping of the applied perturbation because it is possibly diminished by compensatory behavior. The difference between spoken (solid) and heard (transparent) signals shows the mismatch between production and perception. A perturbed signal that equals the baseline mean while the produced signal shows a deviation would indicate perfect compensation. The visible patterns of articulatory behavior over the course of the experiment will be analyzed further below.

IV. STATISTICAL METHODS AND RESULTS

The subsequent statistical examinations aim at capturing three key effects of the present temporal auditory feedback perturbation paradigm extracted from ramp, hold, and aftereffect phases.

First, the ramp phase provides information about the reaction threshold and sensitivity to gradually increased perturbation (Sec. IV A). Second, the hold phase analyses show the directionality and magnitude of differences in hold phase productions relative to baseline productions per segment (CC and V) when maximum perturbation is applied (Sec. IV B 1). Additionally, the reaction magnitude of the whole perturbed segment (CCV and VCC) is set in relation to the applied amount of perturbation (Sec. IV B 2). Last, the aftereffect phase analysis provides the span of trials for which reactions may persist when normal feedback is abruptly restored (Sec. IV C).

Each phase was modeled individually to capture the within-phase behavior. Modeling over phase boundaries (statistically or visually) could distort timepoint specific effects related to the very different perturbation statuses of the trials (e.g., the abrupt transition of maximum perturbation to no perturbation from hold phase to aftereffect phase) and was thereby avoided.

Statistical analyses were conducted with RStudio (RStudio, 2015; R Core Team, 2018) and selected with respect to expected reaction patterns based on the applied perturbation.

A. Ramp phase

In the ramp phase, linearly increasing perturbation was applied. With a possible delay in reaction, caused by the need for a threshold that makes a perturbation (subconsciously) audible, we expected a linear or nonlinear function in production diverging from the baseline mean. For this instance, general additive mixed models (GAMMs) were fitted to the ramp phase. GAMMS account for linear or nonlinear relationships in the data by relying on parametric terms and smooth terms. The smooth terms define the shape of the

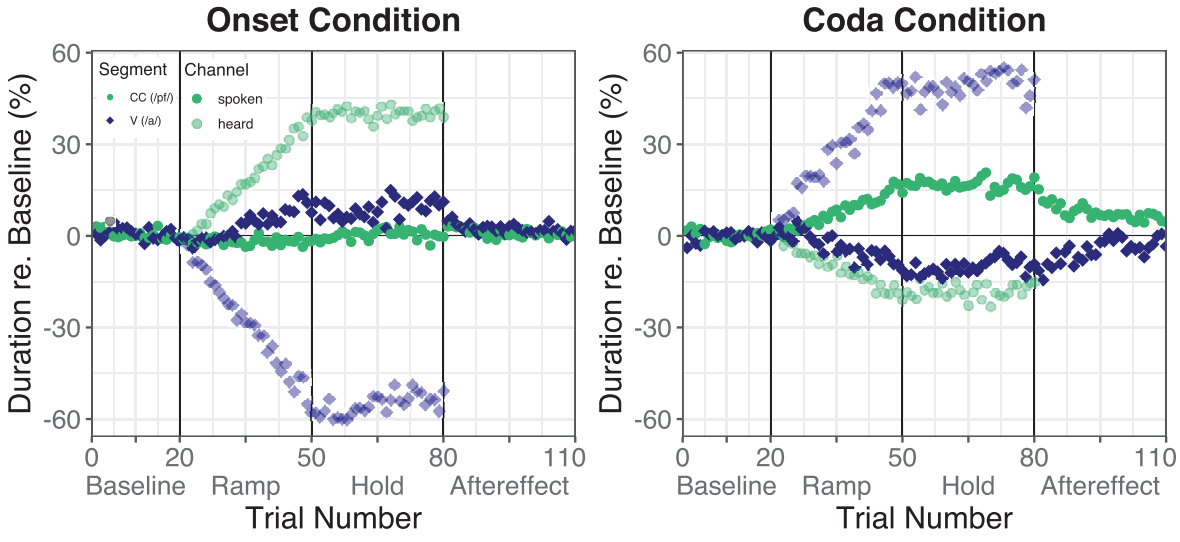

FIG. 2. (Color online) Normalized relative durations averaged over all subjects ($n = 34$ for the onset condition, $n = 33$ for the coda condition) per trial. The vowel /a/ is shown in blue rhombuses and CC /pf/ is shown in green round dots. The spoken signal is shown in solid colors and the perturbed (heard) signal is shown with higher transparency. The left panel visualizes the onset condition and the right panel visualizes the coda condition.

https://doi.org/10.1121/10.0001765

fitted curve by adding up basis functions to a more complex curve until it fits the data properly. Unlike general additive models (GAMs), the mixed design incorporates random effects. Additionally to random slope and random intercept, a random smooth parameter enables the capturing of by-group variation in the nonlinear effects (Sóskuthy, 2017).

With the *R* packages *mgcv* (Wood, 2011, 2017) and *itsadug* (van Rij *et al.*, 2017), one model was fitted per perturbation condition (onset/coda), including both segments of interest (CC/V). The data included trials of the ramp phase (trials 21–50) exclusively. The GAMMS were fitted to normalized relative durations (the outcome variable) with the following terms: segment (V or CC) as a parametric term (average difference in normalized relative duration depending on segment), a smooth term over the trial number (nonlinear effect of the trial number on the normalized relative duration) by segment, and a by-segment factor random smooth nested within subject over trial number with penalty order $m = 1$ (to model inter-speaker variation).

The models were calculated to visualize the significant reaction over time rather than to report *p*-values. Statistical results could summarize comparisons of the means between ramp phase and baseline, which is not necessarily useful when the main interest lies in the point in time (trial number) where reactions start to diverge significantly from the baseline. Visualizations of the models provide the span of the trials with significant effects for each segment (Fig. 3). These indicate how sensitively subjects react to the introduction of perturbation.

In the onset condition, the model suggested a significant deviation from 0 for the vowel around trial number 35 (15 trials after perturbation onset, compression of the perturbed part to ∼61% of its original length) to the end of the ramp phase. No significant effect was found for the CC segment. In the coda condition, vowel durations differed significantly from 0 from trial 33 to the end of the ramp phase (13 trials after perturbation onset, stretching of the perturbed part to ∼133% of its original length), and a significant reaction for the CC segment from trial number 27 to the end of the ramp phase (7 trials after perturbation onset, compressing the perturbed part to ∼83% of its original length). Figure 3 shows the produced differences over the ramp phase and the span of significant deviation from the baseline mean (zero). With a significant effect around the same trial for the vowel in onset and coda conditions, the sensitivity to vowel perturbation seems not to be influenced by the perturbation direction (stretching or compressing) or whether it is the first or second perturbed segment.

B. Hold phase

1. Produced segment durations

The trials of the baseline and hold phase were exposed to a continuous amount of perturbation, either to no perturbation (all baseline trials) or maximum perturbation (all hold trials). Consequently, a systematic effect over time within one of the phases is not assumed. Therefore, linear mixed models were fitted to estimate the differences between baseline and hold phase productions using the packages *lme4* (Bates *et al.*, 2015) and *lmerTest* (Kuznetsova *et al.*, 2017). One model was fitted per perturbation condition (onset/coda), including both segments of

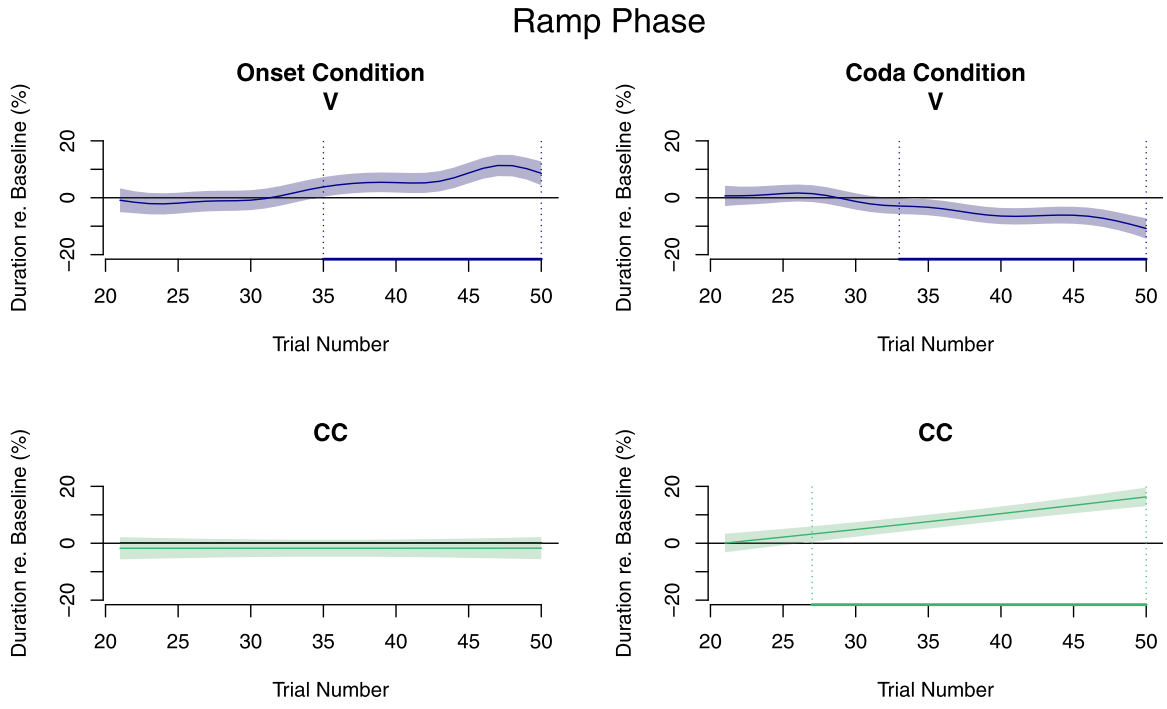

FIG. 3. (Color online) GAMM fits of the ramp phase, including random effects and confidence intervals (95%). The onset condition appears in the left panels (34 subjects) and the Coda condition appears in the right panels (33 subjects). CC fits are shown in green and vowel fits are shown in blue. Dotted vertical lines and thick horizontal lines mark the significance from zero for each sound.

interest (V and CC). The normalized relative durations were modeled as dependent variable with segment (V and CC) and phase (baseline and hold phase) as predictors and an interaction between segment and phase. Random effects included by-subject intercepts and random slopes for phase and segment.

Post hoc pairwise comparisons on significant effects between the phases per segment were performed using the *emmeans* package (Lenth *et al.*, 2018). The significance level was Bonferroni-corrected as we calculated two models for onset and coda conditions ($\alpha = 0.025$). The *post hoc* comparisons for the onset condition returned a significant average lengthening of 8.8% (~11.5 ms) for the vowel /a/ [estimate, 8.76; standard error (SE), 1.59; degrees of freedom (df), 38.78; *t*.ratio, 5.5; $p < 0.025$]. No significant effect was indicated for the CC segment /pf/ [average lengthening of 0.5% (~2 ms; estimate, 0.5; SE, 1.59; df, 38.77; *t*.ratio, 0.317)]. For the coda condition, the model revealed significant effects for the vowel /a/ with an average shortening of 10.3% (~9 ms), which indicated a significant compensatory response (estimate, −10.29; SE, 1.27; df, 42.72; *t*.ratio, −8.1; $p < 0.025$). For the CC segment /pf/, the model indicated a significant compensatory response with an average lengthening of 17.2% (~34 ms) in the hold phase relative to the baseline (estimate, 17.15; SE, 1.27; df, 42.72; *t*.ratio, 13.48; $p < 0.025$). Figure 4 summarizes the durations in the hold phase relative to the baseline mean (zero).

For completeness, linear mixed models with specifications similar to those above were fitted for the single consonants /p/ and /f/. One model was fitted per perturbation condition, comprising both sounds of interest. As previously, *post hoc* testing with a Bonferroni-corrected significance level revealed results for the single sounds. For the onset consonant sequence /pf/ (onset condition), the model reported a nonsignificant average shortening of 2.7% (~3 ms) for C1 /p/ (estimate, −2.72; SE, 1.73; df, 54.88; *t*.ratio, −1.57) and a nonsignificant lengthening of C2 /f/ of 3.8% (~5 ms; estimate, 3.85; SE, 1.73; df, 54.85; *t*.ratio, 2.22). For the coda condition, significant lengthening for both sounds was observed with 18.7% (~15 ms) for C1 /p/ (estimate, 18.71; SE, 2.47; df, 43.59; *t*.ratio, 7.58; $p < 0.025$), and 17.4% (~19 ms) for C2 /f/ (estimate, 17.45; SE, 2.47; df, 43.58; *t*.ratio, 7.07; $p < 0.025$).

Figure 5 visualizes normalized relative durations for the whole CC segment (green dots), C1 (blue squares), and C2 (orange triangles). The spoken signal is shown in solid colors, the perturbed (heard) signal is shown with higher transparency. As a caveat, if the subject adjusted productions for the first part of the perturbation section (first sound onset condition, C1 /p/; coda condition, V /a/), the sound in the middle of the perturbation section (onset condition, C2 /f/; coda condition, C1 /p/) could not be ensured to be always perturbed in the right direction since temporal adjustments altered the fit of the perturbation section (see Fig. 1 for visualization of the fit of the perturbation section). Figure 5 indicates that in the onset condition, both single consonants have been stretched in perturbation (transparent dots, squares, and triangles). In productions, C1 has been rather compensatorily shortened (blue solid squares) while C2 /f/ has been lengthened, indicating a following of the perturbation (orange solid triangles). In the coda condition, C1 /p/ remained mostly unaffected by the perturbation (because both the spoken and the heard signal have approximately the same durations, solid and transparent blue squares), whereas C2 /f/ was compressed (orange transparent triangles). Still, both sounds were lengthened in production, compensating for the duration of the whole CC segment (solid triangles and squares). The observed patterns will be further interpreted in Sec. V (the discussion).

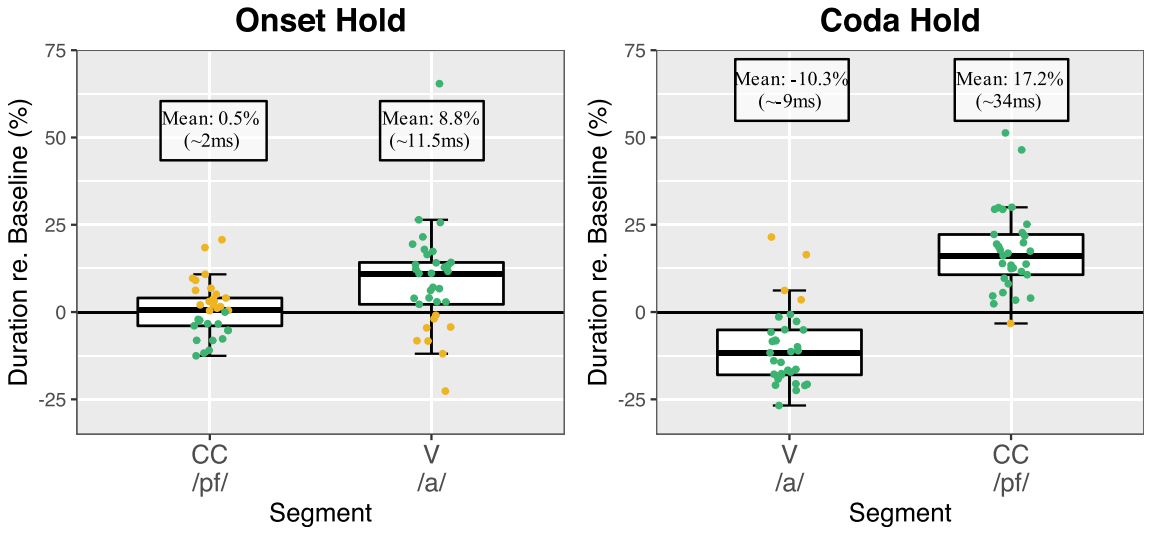

FIG. 4. (Color online) Normalized relative durations in the hold phase relative to the baseline mean (0) for vowel /a/ and CC /pf/ in the onset perturbation condition (34 subjects, left panel) and coda perturbation condition (33 subjects, right panel). Boxes correspond to the first and third quartiles and bars represent the median. Whiskers extend from the hinge to the highest/smallest value but no further than 1.5 interquartile range (IQR). Data beyond the whiskers are outliers. Individual subjects are represented with colored dots where Green dots mark the compensatory behavior and golden dots mark a following of the perturbation direction.

https://doi.org/10.1121/10.0001765

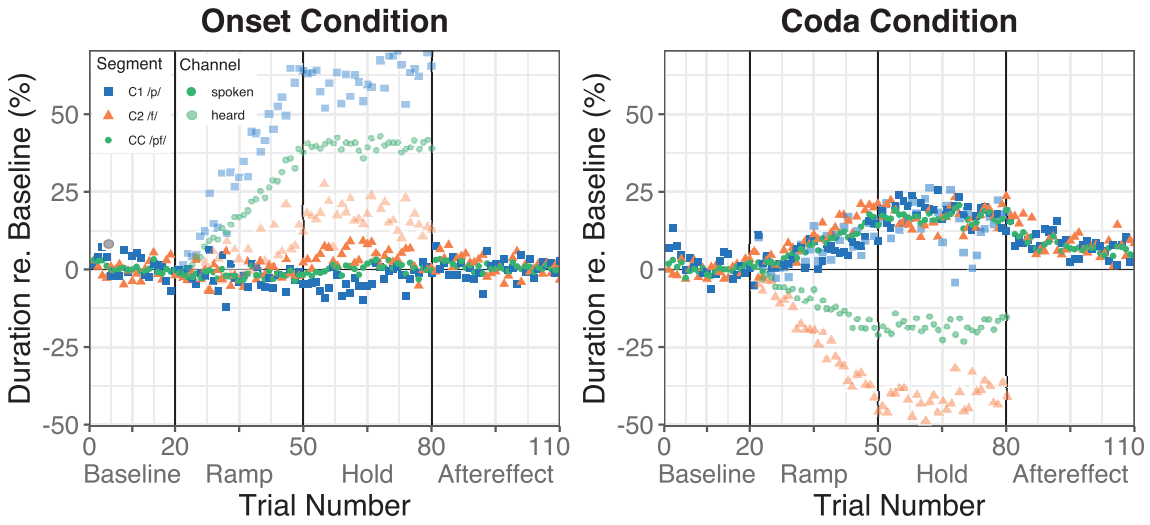

FIG. 5. (Color online) Normalized relative durations averaged over all subjects ($n = 34$ for the onset condition, $n = 33$ for the coda condition) per trial. The CC /pf/ is shown in green round dots, C1 /p/ is shown in blue squares, and C2 /f/ is shown in orange triangles. The spoken signal is shown in solid colors, and the perturbed (heard) signal is shown with higher transparency. The onset perturbation condition is shown in the left panel and the coda perturbation condition is shown in the right panel.

2. Compensation relative to perturbation

The analysis of duration differences between the baseline and hold phase has shown that subjects are capable of compensatory responses for perturbations in the temporal domain in both directions (i.e., shortening of the vowel and lengthening of CC in the coda condition). The compensation values represented the produced duration difference relative to the baseline. To determine how strong this compensation was relative to the applied perturbation, an additional measure was calculated that incorporates the amount of perturbation and takes into account that perturbation is applied to sounds that may already be produced compensatorily. Further, reactions to the whole perturbed sequence (CCV /pfa/, onset condition, and VCC /apf/, coda condition) were taken into consideration. To estimate the relation between applied perturbation and compensation of a segment, absolute sound durations form the bases for the following calculations. These give insight into the strength of the reaction relative to perturbation and allow a comparison between onset and coda perturbations for the whole perturbed sequence (VCC/CCV). To ensure a clean comparison between onset and coda conditions, only subjects with data in both perturbation conditions were included in the following calculations (28 subjects; mean, 23 years old; 23 females).

The point of departure is a two-dimensional coordinate system, wherein the segment durations of the first segment (CC for the onset condition and V for the coda condition) are on the x axis and the durations of the second segment (V for the onset condition and CC for the coda condition) are on the y axis [for visualization, see Figs. 6(A) and 6(B)].

For the following calculations, two signals were considered for each phase, baseline (B) and hold phase (H): the original signal spoken by the subject (1) and the perturbed feedback signal heard by the subject (2). Although there was no perturbation applied in the baseline, a simulation of the signal with perturbation was generated to estimate the maximum perturbation on a signal without reaction (B2*). The durations were referenced to mean baseline productions (B1), hence, B1 is at the zero-crossing for both axes. As before, for the calculation of the baseline mean, the first nine baseline trials were excluded. Examples of the signals can be found in Fig. 1. Figure 1(A) shows the signal of a baseline trial spoken by a subject (B1) and below the simulated perturbation of that signal (B2*). Figure 1(B) shows the production of a hold trial from the same subject (H1) and the perturbed signal of the same trial below (H2).

A *mean perturbation* was calculated from the mean of (simulated) maximum perturbation without compensation in the baseline (Euclidian distance |B1-B2*|, Figs. 6(A) and 6(B), dashed line) and perturbation on a signal that perhaps includes a reaction in the hold phase [Euclidian distance |H1-H2|, Figs. 6(A) and 6(B), dashed line; see Eq. (1)]. Assuming that subjects intuitively aim to match the received auditory feedback with the intended speech sound through compensation, a closer distance between B1 (spoken and heard signals without perturbation) and H2 [heard signal (perturbed auditory feedback) in the hold phase] would mean a stronger compensation. If H2 equals B1, the reaction is interpreted as perfect compensation, meaning that the subject heard the signal he or she intended to speak. The Euclidian distance of |B1-H2| (solid line) was then divided by the *mean perturbation* and scaled to percent values [see Eq. (2)] forming our *compensation* values.

$$\text{mean perturbation} = \frac{|B1 - B2| + |H1 - H2|}{2} \tag{1}$$

$$\text{compensation} = 1 - \left(\frac{|B1 - H2|}{\text{mean perturbation}}\right)100 \tag{2}$$

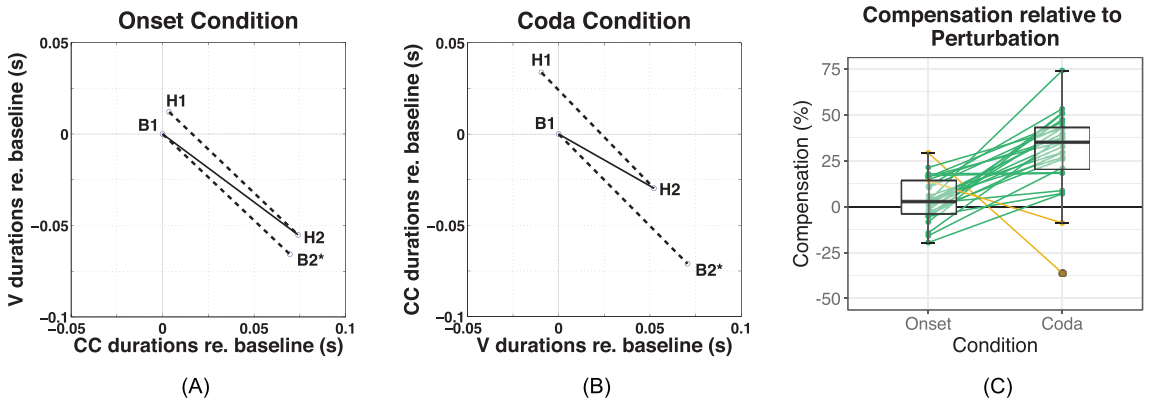

FIG. 6. (Color online) (A) and (B) show mean durations (s) of both segments of interest (V /a/ and CC /pf/) over 28 subjects per perturbation condition relative to the baseline mean (0/0). The first segment of the perturbation section is on the x axis and the second segment of the perturbation section is on the y axis. Points labelled "B" mark baseline durations and "H" marks the hold phase durations. B1 and H1 represent the signal spoken by the subject, B2* and H2 represent the (*simulated) perturbed feedback. (A) shows the onset condition and (B) shows the coda condition. (C) The compensation magnitude relative to perturbation for onset and coda perturbation conditions for 28 subjects. Values incorporate both perturbed segments of interest (V /a/ and CC /pf/). Boxes correspond to the first and third quartiles and bars represent the median. Whiskers extend from the hinge to the highest/smallest value no further than 1.5 IQR. Data beyond the whiskers are outliers. Dots mark individual subjects and are linked by solid lines. Green dots/lines mark those subjects that compensated more in coda than in onset ($n = 26$) and gold dots/lines mark the subjects that compensated more in onset than in coda ($n = 2$).

Based on these calculations, we observed compensation relative to perturbation between −19% and 29% for the onset condition [mean = 4%, standard deviation (sd) = 11.7, median = 3%], and between −36% and 74% (mean = 31%, sd = 21.5, median = 35%) for the coda condition. A negative value results from a following of the perturbation (for at least one of the perturbed segments /a/ or /pf/). A paired t-test was executed to estimate the relation of onset compensation to coda compensation, which turned out to be significant, showing greater compensation in the coda condition [$t = -5.3$, $p < 0.001$, visualized in Fig. 6(C)].

C. Aftereffect phase

The preceding analyses of the hold phase showed temporal adjustments as a reaction to the perturbation for all sounds of interest except for the CC segment /pf/ in the onset condition. The following calculations aimed to examine the stability of the produced compensatory adjustments after perturbation was removed. A persistence of articulatory adjustments into the aftereffect phase could indicate that the underlying motor plan of speech execution experienced a stable realignment in connection with the perceived auditory feedback. For the aftereffect phase, in which auditory feedback was abruptly restored, we expected the behavioral data to show either linear or nonlinear functions peaking off from maximum compensation toward the baseline mean again. To capture these possible patterns, GAMMS were fitted over all trials of the aftereffect phase (trials 81–110).

As previously done for ramp phase examination (see Sec. IV A), one GAMM was fitted per perturbation condition (onset/coda) to normalized relative durations (the outcome variable) with the following terms: segment (V or CC) as a parametric term (average difference in normalized relative duration depending on segment), a smooth term over trial number (nonlinear effect of trial number on normalized relative duration) by segment, and a by-segment factor random smooth nested within subject over trial number with penalty order $m = 1$ (to model inter-speaker variation).

The model for the onset condition suggested no significant effect in the aftereffect phase for either the V or CC segment (which was expected for the CC segment because no significant effect was shown during the hold phase). For the coda condition, the model suggested a persistent significant reaction for the vowel until trial 93 and for the CC segment until trial 108, the latter comprising almost the whole aftereffect phase. Hence, persistent articulatory adjustments were shown for both sounds of the coda condition. The GAMM fits are visualized in Fig. 7.

V. DISCUSSION

The data reported in the current study reveal sensitivity to real-time temporal auditory feedback perturbation. Subjects were found to mainly compensate in the opposite direction to the applied shift for the vowel /a/ in both perturbation conditions (onset condition, /pfa/; coda condition, /apf/) for the CC segment /pf/ in the coda condition but not for the CC segment in the onset condition (which will be discussed further below). With a significant effect around the same trial during the ramp phase for the vowel in both conditions, the sensitivity to vowel perturbation seems not to be influenced by the perturbation direction (stretching or compressing) or whether it is the first segment (coda condition) or second segment (onset condition) of the perturbed section.

A. Adaptation and reactive feedback control

In the coda perturbation condition, articulatory adjustments were found to persist significantly for several trials after perturbation was removed for both perturbed segments CC and V. This pattern indicates a fine-tuning of the

https://doi.org/10.1121/10.0001765

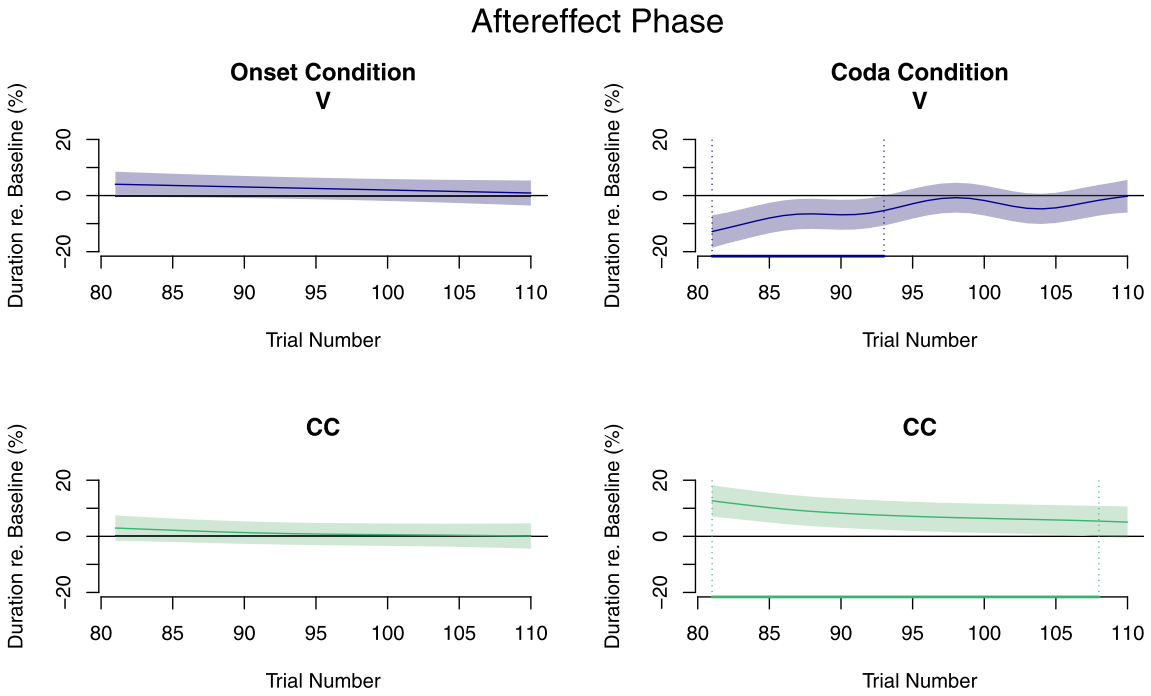

FIG. 7. (Color online) GAMM fits of the aftereffect phase, including random effects and confidence intervals (95%). The onset condition appears in the left panels (34 subjects) and the coda condition appears in the right panels (33 subjects). The CC fits are shown in green and the vowel fits are shown in blue. Dotted vertical lines and thick horizontal lines mark the significant deviation from zero for each sound.

underlying motor plan for the temporal features of the produced speech sounds (adaptation). However, for the vowel in the onset perturbation condition, there was significant compensation during maximum perturbation (hold phase) but no persistent temporal adjustment after normal feedback was restored (aftereffect phase). This effect requires further explanation since we argue that online compensation to perturbed sound duration is not possible: Local adjustments to altered sound durations cannot be processed and executed instantly within the same trial because the duration of a sound is not determinable until it has been entirely perceived. However, the lengthening of the vowel in the CCV condition might not only result from the perturbatory compression of the vowel itself but could also be caused partly by the perturbatory stretching of the onset segment CC.

This leads us to a general remark about the processing possibilities in the first and second halves of the perturbation section: Recall (e.g., from Fig. 1) that the total duration of the perturbation section was on the order of up to 300 ms. Thus, the second half (where perturbatory compression is applied) is about 150 ms from the overall onset of perturbation. Based on what is known about the latency of responses to sudden formant and pitch perturbations, it is possible that the subject response in the second half of our perturbation section is not just compensation for this perturbation but also an online reaction to what has occurred in the first part of the perturbation phase.

The lengthening of the vowel in production might have been a within-trial feedback reaction to the stretched percept of the preceding CC segment with the aim of keeping the relation between CC and V more constant. Contrarily, the timing relations in production between V and CC for the VCC sequence in the coda condition diverge with increased perturbation. The hypothesized reactive feedback control pattern in the onset condition is reminiscent of the findings of Cai *et al.* (2011). They confused the subjects' expectations about the extent of a segment by altering its temporal midpoint (spectral target) but kept the total sound duration constant. Their subjects delayed following productions in the utterance when the perturbed target was decelerated but showed no significant reaction to the acceleration of the spectral target in perturbation.

The more constant temporal relationship between the onset CC segment and V in production indicates greater stability in CCV timing patterns than in VCC sequences. A more stable timing relation in CCV might also be slower to update persistently. Further support for this assumption can be derived from a modeling study by Nam and Saltzman (2003): In modeling the coupling relations of CCV and VCC, they added noise to the coupling potential function, simulating trial-to-trial variability or changes in the speaking rate. They demonstrated that the coupled oscillator model can account for greater stability and different relative timing for onsets in CCV sequences compared to codas in VCC sequences when variability is increased. If we consider this interference to the system as a form of perturbation, then their study found, in the gestural domain, similar effects to the acoustic results of the present study regarding onset stability. Consequently, there might have been some update of temporal vowel representation in the CCV condition, but this was clearly less stable than the update for the perturbed segments in the coda condition. The persistent

adjustments for both of the coda segments indicate predominately adaptive behavior.

Adaptation effects have been shown before for spectral parameters of speech, e.g., in formant or pitch manipulations (Jones and Munhall, 2000, 2002; Purcell and Munhall, 2006b; Villacorta *et al.*, 2007) or for alterations of CoG in fricatives (Shiller *et al.*, 2009). In perturbation of temporal parameters of speech, Mitsuya *et al.* (2014) reported bidirectional adaptation effects for temporally altered VOT of word-initial plosives. Their study showed for the first time that temporal properties of speech are influenced by auditory feedback and can be compensated for in a predictive manner, albeit not in real-time. Very recently, Floegel *et al.* (2020) showed adaptive shortening for stretched vowels or fricatives in real-time.

With the adaptation paradigm of the current study, it was for the first time possible to elicit bidirectional reaction patterns, viz., lengthening and shortening of segments in multisyllabic speech as a compensatory reaction to a real-time perturbation. Further, the data of the current study indicate that the nature of the reaction to temporal perturbation is affected by the syllable position, which has not been found before. Unlike Mitsuya *et al.* (2014), the current study did not reveal compensatory adaptation of timing properties in the onset position. However, the effects of both studies should be compared with caution since Mitsuya *et al.* (2014) manipulated a part of a sound (VOT) rather than the total duration, with the manipulated part, moreover, functioning as a distinctive phonological cue. The unraveling of the manipulated CC /pf/ onset segments in the current study indicated similar effects to those of Mitsuya *et al.* (2014) in the sense that subjects showed a certain amount of compensatory shortening for the initial plosive C1 /p/ (Fig. 5, blue solid squares). Then again, subjects rather followed the perturbation direction in production by lengthening C2 /f/ (Fig. 5, orange solid triangles). Taken together, this resulted in an (almost) equal duration of the whole CC onset segment throughout the experiment (Fig. 5, green solid dots). This indicates that it is, in principle, possible to elicit some temporal articulatory adjustments in the onset of a syllable (since there is a tendency for compensation of the first, leftmost consonant /p/), but in complex onsets, the timing of the whole onset segment seems to be of higher motor priority. In contrast, both consonant segments of the coda condition showed an equally strong (compensatory) response in the same direction, resulting in an overall lengthened CC coda segment.

However, with an interaction between adaptation and within-trial reactive feedback control due to the stretching-compressing paradigm of the current study, it could also be the case that subjects lengthen /f/ in the onset condition as a reaction to the previous longer perceived /p/ (Fig. 5, left panel, transparent blue rhombuses). This applies also to the middle sound in the coda condition: The /p/ was mostly lengthened in production even though (or due to the fact that) it was not much affected by the perturbation. The lengthening could have been a reaction to the longer percept of the preceding vowel. Nevertheless, even after taking these potential interferences into account, there still remain greater articulatory adjustments for the coda perturbation than for the onset perturbation. Thus, the compensatory behavior persists for the first perturbed sound of the coda condition (vowel /a/) but does not persist for the first sound of the onset condition (consonant /p/), which underlines the different nature of the compensatory behavior in onset vs coda perturbation.

Taking stock up to this point, we would contend that the shortening in production of the vowel in the coda condition must be an adaptive response (even in the hold phase) since this sound is located in the first half of the perturbation section before a reactive response seems plausible. The response for the coda CC segment could have some reactive component (as just discussed), but given the clear adaptive response for V in the coda and the clear aftereffect for CC, a strong adaptive component seems very likely.

For the onset condition, there is no unequivocal evidence of adaptive effects, i.e., very little happens to the segments located in the first part of the perturbation section in the hold phase, and there are no aftereffects for any segments. So even if it is not conclusively demonstrable just with the data of this experiment, it is nonetheless tempting to conclude that the predominant effects in the onset condition are within-trial reactive responses. In short, this leads to our overall conclusion that temporal feedforward representations are much less malleable in the onset.

To examine the interaction between adaptation and within-trial reactive feedback control more precisely, less complex stimuli could be chosen with similar sounds in the onset and coda positions.

B. Sensory interdependence and feedback processing

When comparing onset and coda behavior, it remains a concern that they have been treated differently in perturbation. Whereas in the onset condition the CC segment was mostly stretched, it was mostly compressed in the coda condition (and vice versa for the vowel). Additionally, it can be assumed that different sounds show different sensitivity to perturbed auditory feedback. However, there is, to our knowledge, no systematic prediction about why certain sounds could only show adaptive behavior in one direction (either lengthening or shortening, although there has to be a physiological restriction in shortening), and the perturbation of the same sounds in onset and coda conditions should counterbalance for sound specific behavior.

The current study reported compensation magnitudes relative to the applied perturbation of around 4% for onset + vowel perturbation and 31% for vowel + coda perturbation (Sec. IV B 2). The compensation to onset perturbation was smaller overall than for coda perturbation due to the nonsignificant reaction of the CC onset cluster. In both cases, compensation remains incomplete, as previously found for spectral auditory feedback perturbations with compensation values of 25%–30% (Max *et al.*, 2003; Purcell and Munhall, 2006a; MacDonald *et al.*, 2010;

https://doi.org/10.1121/10.0001765

Mitsuya *et al.*, 2011). Partial compensation for auditory shifts has mainly been attributed to the contribution of somatosensory feedback to speech production. When the auditory feedback is altered, the somatosensory feedback remains unchanged. Once articulation changes in the course of compensation for the auditory discrepancy between target and feedback, the mismatch in the auditory domain might decrease. Concurrently, however, the mismatch between the somatosensory target and somatosensory feedback increases.

Research on the interaction between somatosensory and auditory feedback has largely agreed on the latter's predominance ontogenetically with an earlier establishment of auditory targets over somatosensory targets (Guenther, 2006). Later on, adult speakers seem to establish an individual preference about the weighting of the different sensory feedback channels in speech production (Lametti *et al.*, 2012). However, when a mismatch between one sensory reference and the received feedback is introduced (e.g., an auditory feedback perturbation), then not only individual preferences but also the time of exposure and the magnitude of the feedback shift can evoke a dominance of one feedback domain over the other (Purcell and Munhall, 2006b,a; Katseff *et al.*, 2012). Investigations on articulatory initiation have shown that speakers adjust articulator posture before the actual initiation of the utterance, providing earlier access to somatosensory information well before auditory information can be received (Kawamoto *et al.*, 2008; Tilsen, 2016; Krause and Kawamoto, 2019). Additionally, auditory information naturally becomes perceivable later than somatosensory information. In onsets, auditory feedback cannot provide predictions about relative timing within a syllable, unlike the case for codas in which information about onset and vowel duration has already been auditorily received.

Taking this into consideration, we speculate that there is not only an individual preference in sensory reliance but, more intriguingly, also a different weighting in the interplay between somatosensory and auditory feedback with respect to the prosodic position within the syllable. A greater reliance on somatosensory feedback of onsets could explain their greater resistance to updating the motor plan when (only) auditory feedback is perturbed. This idea is reinforced by simulations on stuttering. Civier *et al.* (2010) found that an overreliance on auditory feedback leads to syllable repetitions in onsets, suggesting that people who stutter show impaired read-out of feedforward control and use auditory feedback to a greater extent than fluent speakers.

However, an overreliance on somatosensory feedback in onsets seems to be of higher importance for speech timing than for spectral speech targets: The study by Shiller *et al.* (2009) showed that spectral perturbation of the CoG of /s/ and /ʃ/ in onset positions led to compensatory responses, indicating that auditory feedback seems to play a role for adjustments of *spectral* properties of speech sounds in onsets.

Evidence for different processing of temporal and spectral auditory speech information comes from the study by Floegel *et al.* (2020). They tested lateralization of hemispheric activation during the dichotic presentation of spectrally or temporally altered stimuli. In neuroanatomical approaches to modeling speech production, the left hemisphere is suggested to predominately host feedforward specifications, whereas the right hemisphere processes auditory feedback (Tourville and Guenther, 2011). In auditory perception, however, spectral features have been found to be processed with right-lateralization while temporal features are rather processed with left-lateralization (Flinker *et al.*, 2019). As the first study that combined both spectral and temporal auditory feedback perturbation with fMRI, Floegel *et al.* (2020) were able to show that both hemispheres are involved in auditory feedback control with a right-lateralization during spectral perturbations and a left-lateralization during temporal perturbations. The localization of both temporal processing and speech motor programs in the left-hemisphere could underline our assumption that critical temporal information for speech flow might be more entrenched in the motor plan.

As a further interim summary before moving on, let us note here that the preceding argumentation addresses both feedback and feedforward mechanisms with the suggestion that (1) speakers do not use auditory feedback for the timing of onsets to the same degree as they do for codas, and (2) the mismatch is (subconsciously) detected, but the motor system is not capable of ultimately updating the putatively very stable onset timing patterns in production within the time-span of the experiment.

C. Nature of timing mechanisms in speech and nonspeech

Coupling the idea that timing mechanisms for onsets and codas rely to a different extent on auditory feedback control with research on predictive timing, we can draw parallels to other nonspeech timing mechanisms that demand prediction. Previous research outlined a distinction between at least two timing mechanisms: relative/event-based timing, which occurs relative to a predicted rhythmic event such as a musical beat, and absolute/duration-based timing, which is established on the absolute estimation of temporal intervals (Grube *et al.*, 2010; Teki *et al.*, 2011; Arnal and Giraud, 2012; Teki *et al.*, 2012; Grahn and Watson, 2013). Recent neuroscientific research suggests that predictive timing in both music and speech perception may be underpinned by similar mechanisms, whereby recurrences of syllable onsets are comparable to beats in music even if the former occur only at quasi-periodic intervals in speech (Nozaradan *et al.*, 2012; Peelle and Davis, 2012). Further, there have also been indications that forward prediction in music and language may draw upon common timing mechanisms (Iversen *et al.*, 2009; Tierney and Kraus, 2014).

We consider that both timing mechanisms, event-based and duration-based timing, might be involved when making temporal predictions in complex auditory stimuli such as speech. Accordingly, onset timing might likely be driven by event-based timing mechanisms, whereby onset productions

aim at ensuring a continuous speech flow. On the other hand, the nucleus and coda of syllables contribute less to syllable timing and might rather be predicted and executed with underlying duration-based timing mechanisms within a word or syllable time frame. It was found that event- and duration-based timing mechanisms are also associated with different brain regions. Teki *et al.* (2011) found a higher activation in a striato-thalamocortical network during event-based timing, comprising *inter alia* the supplementary motor area and premotor cortex. Additionally, significant activations in an olivocerebellar network comprising the inferior olive, vermis, and deep cerebellar nuclei, including the dentate nucleus, were shown for duration-based timing. The premotor cortex and supplementary motor area were found to be crucially relevant for the planning of internally generated complex motor movements within a precise timing plan rather than relying on sensory information (Roland *et al.*, 1980; Gerloff *et al.*, 1997). A classification of onset timing as an event-based timing mechanism could explain the greater resistance of onsets to temporally perturbed auditory feedback due to a greater reliance on established internal predictive models firmly anchored in the motor plan. This assumption is partially in line with previous research by Kotz and Schwartze (2010), who attributed the planning of temporal structure to the pre-supplementary motor area and basal ganglia. Hereby the cerebellum serves as a pacemaker for basic temporal structure, constituting a grid for the temporal alignment of memory representations.

With the findings of the current study, we assume that those planning mechanisms play a role for timing functionality in speech production dependent on syllable structure. More support for this hypothesis comes from research on people who stutter. It was shown that people who stutter show different activity compared to fluent speakers in brain regions that are involved in timing mechanisms, namely, the basal ganglia-thalamocortical circuit and the cerebellum (Brown *et al.*, 2005; Watkins *et al.*, 2007; Chang and Zhu, 2013). Hence, people who stutter show connectivity differences compared to fluent speakers in neural networks, which are associated with self-initiated movement and internal generation of rhythm (Chang and Zhu, 2013). In stuttering, deficits occur not only in onsets of speech syllables; timing deficits have also been reported in nonverbal beat-alignment tasks that demand event-based timing predictions (Falk *et al.*, 2015). Additionally, people who stutter improve speaking fluency when their speech is accompanied by an external paced beat like a metronome. These observations strengthen the assumption that onsets might be associated with event-based predictive timing mechanisms while codas rather follow principles of duration-based timing mechanisms with the latter being influenced to a greater extent by auditory feedback information.

Certainly, these assumptions need further verification, for example, by testing the brain regions involved in both discussed predictive timing mechanisms with respect to their activity while producing and perceiving speech with special attention to syllable structure.

D. Models of speech production

The compensatory responses in the current study indicated a crucial contribution of auditory feedback to timing mechanisms in speech on both the control and planning levels. While the compensatory behavior in the coda perturbation condition indicated adaptation of temporal properties on a within-phoneme level, the reactions to onset perturbation rather suggested reactive online compensation for perturbed timing relations on a within-syllable level. Further, the results underline the fact that representations of speech segments must comprise information about segment duration that can be adjusted dynamically and updated when needed.

In attempts to interpret these findings within the scope of speech production models, there is, to our knowledge, no model which can comprehensively account for these results: Adaptation and reactive control of speech timing through auditory feedback need a specification of timing relations that is sensitive to syllable position but includes the contribution of auditory feedback on the planning and control levels. While adaptation to spectral perturbations of speech is well explainable with several models that include a representation of spectral state variables and feedback mechanisms, we would like to contend that duration as a property of speech sounds needs to be treated and modeled differently: State variables, such as frequency, intensity or pitch, evolve in time. Duration, however, marks the extent of this evolution over time (Tilsen, 2019).

As one of the most comprehensive approaches to modeling speech production, DIVA assumes auditory speech targets that consist of time-varying spectral properties. With the data of the current study, it seems likely that the *extent* of those spectral features over time (duration) must also be inherent to the motor plan and can be established and updated through auditory feedback.

The findings of the current study support once more the motivation for modeling timing aspects in speech production with an involvement of sensory feedback on control and planning levels.

E. Individual behavior

As a final point, note that in the current study we presented data mostly summarized over all subjects with graphical representation of single subjects [in Figs. 4 and 6(C)]. Lately, a number of studies reported systematic differences in reaction to perturbed auditory feedback on the subject level. While the majority of subjects compensated for an applied shift (as summarized in many studies), there are quite a few reports of subjects who followed the direction of the perturbation (see, e.g., Burnett *et al.*, 1998; Hain *et al.*, 2000, for pitch perturbation; Purcell and Munhall, 2006b, for formant perturbation; Klein *et al.*, 2019, for fricative perturbation). Further subjects were reported not to show a consistent reaction at all, varying between following and compensatory responses between adjacent trials (Behroozmand *et al.*, 2012) or shift directions (Klein *et al.*,

2019). Varying responses on inter- and intra-subject level were indeed observed in the current study, as for the example marked in Fig. 4 (green dots mark compensatory responses, gold dots mark following responses). Nevertheless, our attempts to group subjects into followers or compensators for the whole study or one perturbation condition did not result in a reasonable grouping or lead to any behaviorally explicable pattern since there were two perturbation conditions (onset and coda perturbations) each consisting of two perturbation directions (stretching and compressing), resulting in four observed segments. On an individual level, some subjects, for example, compensated for three of them but followed for one. Patterns such as these undoubtedly contributed to the high variance in the overall measure of compensation magnitude relative to applied perturbation when summarizing both segments (/a/ and /pf/) per perturbation condition (Sec. IV B 2). We will not explore this further here, but individual differences in compensatory response to temporal perturbations and their origins could be a specific focus of interest and linked to individual rhythmic and temporal discrimination abilities in future investigations.

ACKNOWLEDGMENTS

This work is supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under Grant No. HO 3271/6-1. We thank Sebastian Böhnke and Valeria Trubnikow for their help in running the tests and our participants for taking part in the study. We also thank Michele Gubian and Stefano Coretta for their help with GAMMS and three anonymous reviewers for their most valuable comments on a previous version of the manuscript. Lastly, we would like to thank Shanqing Cai and colleagues for making the AUDAPTER software package available.

[1]Note that Saltzman and Munhall (1989) did, in fact, envisage the possibility of feedback from the interarticulator to the intergestural level. See Shaw and Chen (2019) for further examination of the viability of the feedforward assumption in current versions of the model.

Arnal, L. H., and Giraud, A.-L. (**2012**). "Cortical oscillations and sensory predictions," Trends Cognit Sci **16**, 390–398.
Bates, D., Maechler, M., Biolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effect models using lme4," J. Statist. Software **67**(1), 1–48.
Behroozmand, R., Korzyukov, O., Sattler, L., and Larson, C. R. (**2012**). "Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control," J. Acoust. Soc. Am. **132**, 2468–2477.
Boersma, P., and Weenink, D. (**1999**). "Praat, a system for doing phonetics by computer (version 5.3.78) [computer program], http://www.praat.org (Last viewed August 16, 2020).
Browman, C. P., and Goldstein, L. M. (**2000**). "Competing constraints on intergestural coordination and self-organization of phonological structures," Les Cahiers de l'ICP, Bull. Commun. Parlée **5**, 25–34.
Brown, S., Ingham, R. J., Ingham, J. C., Laird, A. R., and Fox, P. T. (**2005**). "Stuttered and fluent speech production: An ale meta-analysis of functional neuroimaging studies," Hum. Brain Mapp. **25**, 105–117.
Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice F0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153–3161.
Byrd, D. (**1996**). "Influences on articulatory timing in consonant sequences," J. Phonet. **24**, 209–244.
Cai, S. (**2014**). *A manual of Audapter. Version 2.1.012* (Speech Laboratory, Department of Speech, Language and Hearing Sciences, Sargent College of Health and Rehabilitation Sciences, Boston University, Boston).
Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2008**). "A system for online dynamic perturbation of formant trajectories and results from perturbations of the mandarin triphthong/iau," in *Proceedings of the 8th ISSP*, pp. 65–68.
Cai, S., Ghosh, S. S., Guenther, F. H., and 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," J. Neurosci. **31**, 16483–16490.
Casserly, E. D. (**2011**). "Speaker compensation for local perturbation of fricative acoustic feedback," J. Acoust. Soc. Am. **129**, 2181–2190.
Caudrelier, T., Perrier, P., Schwartz, J.-L., and Rochet-Capellan, A. (**2016**). "Does auditory-motor learning of speech transfer from the CV syllable to the CVCV word?," in *17th Annual Conference of the International Speech Communication Association (Interspeech 2016)* (San Francisco, CA).
Chang, S.-E., and Zhu, D. C. (**2013**). "Neural network connectivity differences in children who stutter," Brain **136**, 3709–3726.
Chen, Z., Liu, P., Jones, J., Huang, D., and Liu, H. (**2010**). "Sex-related differences in vocal responses to pitch feedback perturbations during sustained vocalization," J. Acoust. Soc. Am. **128**, EL355–EL360.
Civier, O., Tasko, S. M., and Guenther, F. H. (**2010**). "Overreliance on auditory feedback may lead to sound/syllable repetitions: Simulations of stuttering and fluency-inducing conditions with a neural model of speech production," J. Fluency Disord. **35**, 246–279.
Debrabant, J., Gheysen, F., Vingerhoets, G., and Van Waelvelde, H. (**2012**). "Age-related differences in predictive response timing in children: Evidence from regularly relative to irregularly paced reaction time performance," Hum. Mov. Sci. **31**, 801–810.
Donath, T. M., Natke, U., and Kalveram, K. T. (**2002**). "Effects of frequency-shifted auditory feedback on voice F_0 contours in syllables," J. Acoust. Soc. Am. **111**, 357–366.
Etchell, A. C., Johnson, B. W., and Sowman, P. F. (**2014**). "Behavioral and multimodal neuroimaging evidence for a deficit in brain timing networks in stuttering: A hypothesis and theory," Front. Hum. Neurosci. **8**, 1–10.
Falk, S., Müller, T., and Dalla Bella, S. (**2015**). "Non-verbal sensorimotor timing deficits in children and adolescents who stutter," Front. Psychol. **6**, 1–12.
Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., and Poeppel, D. (**2019**). "Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries," Nat. Hum. Behav. **3**, 393–405.
Floegel, M., Fuchs, S., and Kell, C. A. (**2020**). "Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control," Nat. Commun. **11:2839**, 1–12.
Gerloff, C., Corwell, B., Chen, R., Hallett, M., and Cohen, L. G. (**1997**). "Stimulation over the human supplementary motor area interferes with the organization of future elements in complex motor sequences," Brain: J. Neurol. **120**, 1587–1602.
Goldstein, L., Nam, H., Saltzman, E., and Chitoran, I. (**2009**). "Coupled oscillator planning model of speech timing and syllable structure," in *Frontiers in Phonetics and Speech Science*, pp. 239–249.
Goldstein, L., and Pouplier, M. (**2014**). "The temporal organization of speech," in *The Oxford Handbook of Language Production*, edited by M. Goldrick, V. Ferreira, and M. Miozzo (Oxford University Press, New York), pp. 210–227.
Grahn, J. A., and Watson, S. L. (**2013**). "Perspectives on rhythm processing in motor regions of the brain," Music Ther. Perspect. **31**, 25–30.
Grube, M., Cooper, F., Chinnery, P., and Griffiths, T. (**2010**). "Dissociation of duration-based and beat-based auditory timing in cerebellar degeneration," Proc. Natl. Acad. Sci. U.S.A. **107**, 11597–11601.
Guenther, F. H. (**2006**). "Cortical interactions underlying the production of speech sounds," J. Commun. Disord. **39**, 350–365.
Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain Lang. **96**, 280–301.
Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., and Kenney, M. K. (**2000**). "Instructing subjects to make a voluntary response reveals

the presence of two components to the audio-vocal reflex," Exp. Brain Res. **130**, 133–141.
Houde, J., and Nagarajan, S. (**2011**). "Speech production as state feedback control," Front. Hum. Neurosci. **5**, 1–14.
Houde, J. F., and Chang, E. F. (**2015**). "The cortical computations underlying feedback control in vocal production," Curr. Opin. Neurobiol. **33**, 174–181.
Houde, J. F., and Jordan, M. I. (**1998**). "Sensorimotor adaptation in speech production," Science **279**, 1213–1216.
Houde, J. F., and Jordan, M. I. (**2002**). "Sensorimotor adaptation of speech I: Compensation and adaptation," J. Speech, Lang. Hear. Res. **45**, 295–310.
Houde, J. F., Niziolek, C., Kort, N., Agnew, Z., and Nagarajan, S. S. (**2014**). "Simulating a state feedback model of speaking," in *10th International Seminar on Speech Production*, pp. 202–205.
Hubbard, C. P. (**1998**). "Stuttering, stressed syllables, and word onsets," J. Speech, Lang. Hear. Res. **41**, 802–808.
Iversen, J. R., Repp, B. H., and Patel, A. D. (**2009**). "Top-down control of rhythm perception modulates early auditory responses," Annals New York Acad. Sci. **1169**, 58–73.
Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.
Jones, J. A., and Munhall, K. G. (**2002**). "The role of auditory feedback during phonation: Studies of mandarin tone production," J. Phonet. **30**, 303–320.
Katseff, S., Houde, J., and Johnson, K. (**2012**). "Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback?," Lang. Speech **55**, 295–308.
Kawamoto, A. H., Liu, Q., Mura, K., and Sanchez, A. (**2008**). "Articulatory preparation in the delayed naming task," J. Mem. Lang. **58**, 347–365.
Klein, E., Brunner, J., and Hoole, P. (**2019**). "The relevance of auditory feedback for consonant production: The case of fricatives," J. Phonet. **77**, 100931.
Kotz, S. A., and Schwartze, M. (**2010**). "Cortical speech processing unplugged: A timely subcortico-cortical framework," Trends Cognit. Sci. **14**, 392–399.
Krause, P. A., and Kawamoto, A. H. (**2019**). "Anticipatory mechanisms influence articulation in the form preparation task," J. Exp. Psychol.: Hum. Percept. Perform. **45**, 319–335.
Kröger, B. J., Kannampuzha, J., and Neuschaefer-Rube, C. (**2009**). "Towards a neurocomputational model of speech production and perception," Speech Commun. **51**, 793–809.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "Lmertest package: Tests in linear mixed effects models," J. Statist. Software **82**(13), 1–26.
Lametti, D. R., Nasir, S. M., and Ostry, D. J. (**2012**). "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback," J. Neurosci. **32**, 9351–9358.
Lenth, R., Singman, H., Love, J., Buerkner, P., and Herve, M. (**2018**). "Emmeans: Estimated marginal means, aka least-squares means," available at https://cran.R-project.Org/package=emmeans (Last viewed August 16, 2020).
MacDonald, E. N., Goldberg, R., and Munhall, K. G. (**2010**). "Compensations in response to real-time formant perturbations of different magnitudes," J. Acoust. Soc. Am. **127**, 1059–1068.
MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "Probing the independence of formant control using altered auditory feedback," J. Acoust. Soc. Am. **129**, 955–965.
Max, L., and Gracco, V. L. (**2005**). "Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter," J. Speech, Lang. Hear. Res. **48**, 524–542.
Max, L., Wallace, M. E., and Vincent, I. (**2003**). "Sensorimotor adaptation to auditory perturbations during speech: Acoustic and kinematic experiments," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Futurgraphic Barcelona, Spain), pp. 1053–1056.
Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (**2014**). "Temporal control and compensation for perturbed voicing feedback," J. Acoust. Soc. Am. **135**, 2986–2994.
Mitsuya, T., MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (**2011**). "A cross-language study of compensation in response to real-time formant perturbation," J. Acoust. Soc. Am. **130**, 2978–2986.
Nam, H., Goldstein, L., and Saltzman, E. (**2009**). "Self-organization of syllable structure: A coupled oscillator model," in *Approaches to Phonological Complexity*, edited by F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé (de Gruyter, Berlin), pp. 299–328.
Nam, H., and Saltzman, E. (**2003**). "A competitive, coupled oscillator model of syllable structure," in *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2253–2256.
Niziolek, C. A., and Guenther, F. H. (**2013**). "Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations," J. Neurosci. **33**, 12090–12098.
Nozaradan, S., Peretz, I., and Mouraux, A. (**2012**). "Selective neuronal entrainment to the beat and meter embedded in a musical rhythm," J. Neurosci. **32**, 17572–17581.
Parrell, B., Lammert, A. C., Ciccarelli, G., and Quatieri, T. F. (**2019a**). "Current models of speech motor control: A control-theoretic overview of architectures and properties," J. Acoust. Soc. Am. **145**, 1456–1481.
Parrell, B., Ramanarayanan, V., Nagarajan, S., and Houde, J. (**2019b**). "The facts model of speech motor control: Fusing state estimation and task-based control," PLoS Comput. Biol. **15**, e1007321.
Parrell, B., Ramanarayanan, V., Nagarajan, S. S., and Houde, J. F. (**2018**). "Facts: A hierarchical task-based control model of speech incorporating sensory feedback," in *Interspeech 2018*, pp. 1497–1501.
Patel, R., Niziolek, C., Reilly, K., and Guenther, F. H. (**2011**). "Prosodic adaptations to pitch perturbation in running speech," J. Speech, Lang. Hear. Res. **54**, 1051–1059.
Patri, J.-F., Diard, J., and Perrier, P. (**2019**). "Modeling sensory preference in speech motor planning: A Bayesian modeling framework," Front. Psychol. **10**, 1–14.
Patri, J.-F., Perrier, P., Schwartz, J.-L., and Diard, J. (**2018**). "What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework," PLoS Comput. Biol. **14**, e1005942.
Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Front. Psychol. **3**, 1–17.
Purcell, D. W., and Munhall, K. G. (**2006a**). "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am. **120**, 966–977.
Purcell, D. W., and Munhall, K. G. (**2006b**). "Compensation following real-time manipulation of formants in isolated vowels," J. Acoust. Soc. Am. **119**, 2288–2297.
R Core Team (**2018**). "*R*: A language and environment for statistical computing [computer program]" (R Foundation for Statistical Computing, Vienna, Austria).
Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S. S., and Houde, J. F. (**2016**). "A new model of speech motor control based on task dynamics and state feedback," in *Interspeech 2016*, San Francisco, pp. 3564–3568.
Repp, B. H., and Su, Y.-H. (**2013**). "Sensorimotor synchronization: A review of recent research (2006–2012)," Psychon. Bull. Rev. **20**, 403–452.
Roland, P. E., Larsen, B., Lassen, N. A., and Skinhoj, E. (**1980**). "Supplementary motor area and other cortical areas in organization of voluntary movements in man," J. Neurophysiol. **43**, 118–136.
RStudio, T. (**2015**). in "*Rstudio: Integrated development for R*" (RStudio, Inc., Boston, MA).
Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**, 333–382.
Shaw, J. A., and Chen, W.-R. (**2019**). "Spatially conditioned speech timing: Evidence and implications," Front. Psychol. **10**, 1–17.
Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (**2009**). "Perceptual recalibration of speech sounds following speech motor learning," J. Acoust. Soc. Am. **125**, 1103–1113.
Sóskuthy, M. (**2017**). "Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction," arXiv:1703.05339.
Teki, S., Grube, M., and Griffiths, T. D. (**2012**). "A unified model of time perception accounts for duration-based and beat-based timing mechanisms," Front. Integr. Neurosci. **5**, 1–7.
Teki, S., Grube, M., Kumar, S., and Griffiths, T. D. (**2011**). "Distinct neural substrates of duration-based and beat-based auditory timing," J. Neurosci. **31**, 3805–3812.

Tierney, A., and Kraus, N. (**2014**). "Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (path)," Front. Hum. Neurosci. **8:949**, 1–9.
Tilsen, S. (**2016**). "Selection and coordination: The articulatory basis for the emergence of phonological structure," J. Phonet. **55**, 53–77.
Tilsen, S. (**2019**). "Space and time in models of speech rhythm," Ann. N. Y. Acad. Sci. **1453**, 47–66.
Tourville, J. A., Cai, S., and Guenther, F. (**2013**). "Exploring auditory-motor interactions in normal and disordered speech," Proc. Meet. Acoust. **19**, 060180, 1–8.
Tourville, J. A., and Guenther, F. H. (**2011**). "The diva model: A neural theory of speech acquisition and production," Lang. Cognit. Processes **26**, 952–981.
Tourville, J. A., Reilly, K. J., and Guenther, F. H. (**2008**). "Neural mechanisms underlying auditory feedback control of speech," Neuroimage **39**, 1429–1443.
van Rij, J., Wieling, M., Baayen, R., and van Rijn, H. (**2017**). "Itsadug: Interpreting time series and autocorrelated data using GAMMs. *R* package version 2.3," available at https://cran.r-project.org/web/packages/itsadug/index.html.
Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. **122**, 2306–2319.
Watkins, K. E., Smith, S. M., Davis, S., and Howell, P. (**2007**). "Structural and functional abnormalities of the motor system in developmental stuttering," Brain **131**, 50–59.
Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," J. R. Statist. Soc. Ser. B **73**, 3–36.
Wood, S. N. (**2017**). *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC Press, Boca Raton, FL).
Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (**2004**). "Compensation for pitch-shifted auditory feedback during the production of mandarin tone sequences," J. Acoust. Soc. Am. **116**, 1168–1178.