Using real-time temporal auditory feedback perturbation to probe the cognitive representation of speech timing

Miriam Oschkinat

München 2022

Using real-time temporal auditory feedback perturbation to probe the cognitive representation of speech timing

Inaugural-Dissertation

Zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität
München

vorgelegt von

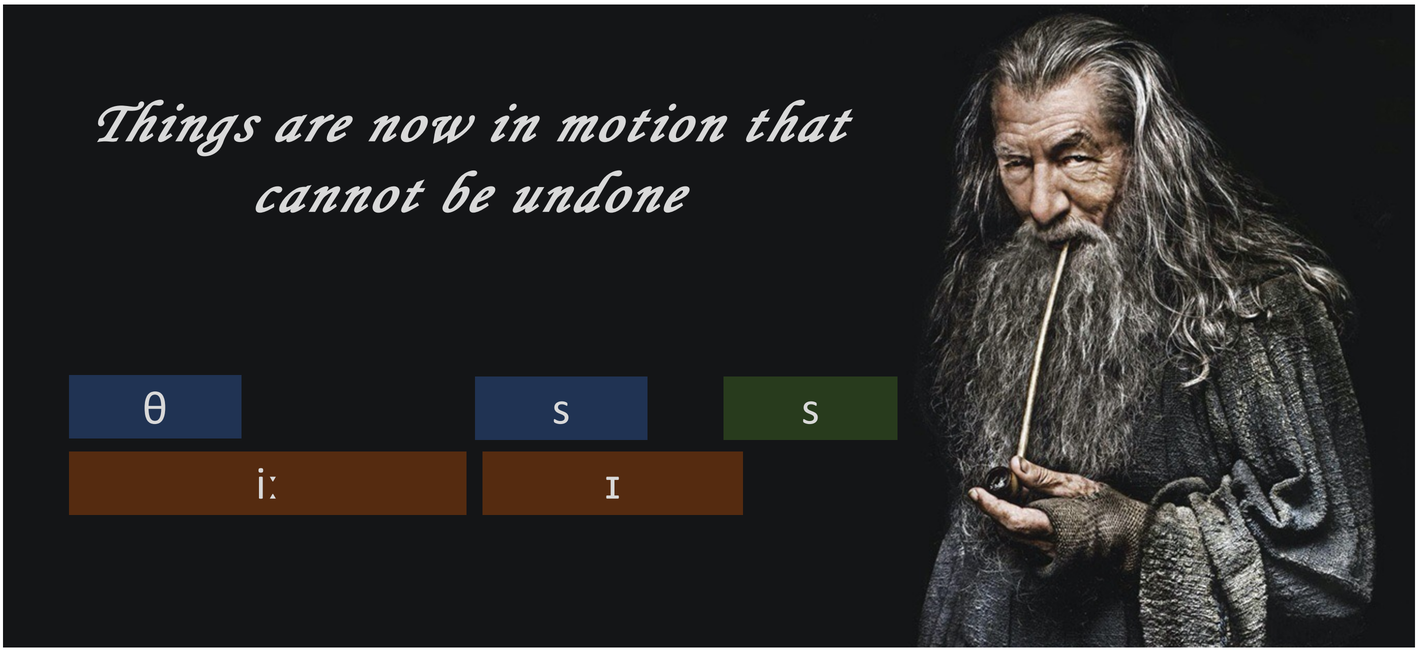
Miriam Oschkinat

München 2022

Erstgutachter: Prof. Dr. Phil Hoole

Zweitgutachterin: Prof. Dr. Eva Reinisch

Tag der mündlichen Prüfung: 12. November 2021



Gandalf in *The Lord of the rings,* by J.R.R. Tolkien
Source image: Wallpaper Access ©2021

To my doubting future self.
You got this.

Content

List of Figures

List of Tables

Chapter 1
General Introduction

As soon as a human opens the mouth, phoneticians and phonologists await with excitement for what they might be told about the mechanisms and consequences of speech production. Human speech requires a large set of physiological and neurological contributors to produce sounds that uniquely serve for verbal communication, each embedded in a language-dependent phonological system that legalizes the coproduction of certain speech elements to larger units of fluent speech. This coproduction requires a temporal organization, which sets the grounding for spectral information to become audible. The generated information from this process is then perceived by a listener and used for the speaker's evaluation of the speech production process.

The temporal organization of fluent speech, and the contribution of auditory feedback to its control and representation, will be the substance of this thesis.

1.1 Speech Production Background

Temporal and spectral properties of speech have been extensively studied over the last decades of phonetic and phonological research. Various research methods have given rise to knowledge that built the basis for speech production theories. These theories have elaborated concepts of how speech properties are established and controlled, foregrounding different contributors and mechanisms dependent on the scientific perspective.

Among the most discussed theories in phonetic research is the *Directions Into Velocities of Articulators* (DIVA) model, which comprehensively describes the establishment and representation of spectral properties of speech via feedforward and sensory feedback mechanisms from a neuro-phonetic perspective (Guenther, 2003; Tourville and Guenther, 2011; Guenther and Vladusich, 2012; Guenther, 2016). According to the DIVA model, speakers establish so-called speech targets of spatio-temporal dimensions for a speech segment (e.g., a sound or a syllable). These targets carry information, for example, on the spectral shape of a sound. In speech production, speakers aim to reach these targets, with the auditory and somatosensory feedback acting as agents for evaluating the fit of the production against the respective target. Speech targets are established in speech acquisition via sensory information, whereby the auditory feedback is used earlier in development than the somatosensory feedback (Guenther, 2016, p.131). Thereby, individual perceptual abilities were shown to shape the establishment of these targets: speakers who can auditorily identify more subtle discrepancies (speakers who have a higher *auditory acuity*) establish smaller target regions and produce more distinct speech (Perkell *et al.*, 2004a; Perkell *et al.*, 2004b; Perkell *et al.*, 2008; Ghosh *et al.*, 2010).

Although speech targets in the DIVA model carry a temporal component, the depiction of this temporal component is not very precise about if and how temporal properties of speech are established and controlled or how they might vary under different prosodic conditions in fluent speech. However, where one theory meets its limitations, another can shine: in contrast to the DIVA model, the Articulatory Phonology/Task-Dynamics framework has formed concepts of temporal coordination in fluent speech.

Therefore, rather than focusing on spectral information, the gestures involved in speech production are the matter of interest (Browman and Goldstein, 1989; Browman and Goldstein, 1992). Gestures are modeled as overlapping structures that are goal-directed and dynamically defined (Saltzman and Munhall, 1989; Saltzman and Byrd, 2000). The coordination of gestures into larger movement patterns varies as a function of context. The specification of this coordination can account for prosodically determined timing mechanisms in fluent speech, such as the c-center effect, and for coarticulation (Browman and Goldstein, 1988). The c-center effect, for example, describes a disparity in temporal organization of segments in onset and coda position within the syllable. Thereby, each onset consonant is coupled with the vowel, while in codas each sound is coupled with its predecessor. The onset therefore has a single midpoint that is timed with the vowel, independent of how complex the onset might be, while in codas, each segment is sequentially coupled (Browman and Goldstein, 1988). Coordinative patterns as such are stored in the *gestural score*, which contains information about the articulators that are active during a task, when and how a constriction should be formed, and how the gestures overlap in time (Browman and Goldstein, 1992).

The use of a dynamical framework for understanding feedforward control is undoubtedly a benefit of the Articulatory Phonology/Task-Dynamics framework compared with the DIVA model when considering fluent speech. Then again, the articulatory approach does not expatiate on how (auditory) feedback mechanisms might contribute to the planning and control of coordinative patterns. If a gesture is activated based on the gestural score, there is no indication on whether it can deviate from its plan in reaction to, for example, erroneous feedback.

The confrontation of articulatorily vs. acoustically orientated theories and methods has triggered a lively debate over the last decades on which are the most prominent components for comprehensively modeling speech production. Many researchers have aimed at merging both approaches or built theories from other perspectives (see Parrell *et al.*, 2019a for an overview of current speech motor control models). One aspect that was implicated in the above introduction of the Articulatory Phonology/Task-Dynamics and DIVA frameworks is the fact that none of those two models, and to our knowledge, none of further existing approaches, makes predictions for the (possible) connection of auditory feedback mechanisms when it comes to planning and control of speech *timing*.

Therefore, the role of the auditory feedback system and its interaction with the feedforward system for planning and flexibly guiding precise timing in fluent speech remains imperative to discover.

1.2 Perturbation

Perturbation experiments have been extensively used to study planning and control mechanisms in speech production. In *auditory* feedback perturbations, the establishment and control of (spectral) speech targets based on feedforward and feedback interaction have been of vital interest. Thereby, a spectral parameter of a speech sound, such as the first formant (F1) frequency, is manipulated in (almost) real-time and fed back to the speaker as auditory feedback. Following the seminal investigation of Houde and Jordan (1998), a large body of research has shown that speakers compensate when spectral features in their auditory feedback are altered (see Caudrelier and Rochet-Capellan, 2019 for an overview). *Compensation* thereby describes a response with adjustments in the opposite direction of the applied shift. When responses indicate that the underlying representations have been changed, speakers are said to *adapt*. Auditory feedback perturbation studies have provided insight into the planning and control of spectral properties of speech and the interaction between feedforward and auditory feedback systems.

A few studies also conducted *articulatory* perturbation experiments, in which one or more of the articulators were perturbed (Folkins and Zimmermann, 1982; Kelso *et al.*, 1984; Baum, 1996; Brunner, 2008; Brunner *et al.*, 2011). For example, in the thesis by Brunner (2008), speakers were provided with an artificial palate and instructed to wear it for two weeks. The study investigated the adaptation to a perturbation of the articulatory space, and the nature of targets in acoustic and articulatory dimensions. The results indicated that both acoustic and articulatory representations must exist, but with a higher priority in producing distinctive acoustic targets unfurling them into a maximized acoustic vowel space. The articulatory representation was suggested to rather serve as a motor function and less for perceptual encoding as speakers compensated instantly for the articulatory perturbation in vowels when masked auditory feedback was provided, further indicating that this representation must be rather flexible. Reorganizations in articulatory strategy under perturbed circumstances aimed at reaching the desired acoustic output with the least effort.

Articulatory and acoustic perturbation paradigms give insight into the goals that speakers aim for and the corresponding motor execution strategies. The *temporal* speech component and its representation, however, has rarely been the focus of these investigations. Articulatory perturbations of exclusively temporal properties seem impossible, as an articulatory perturbation constantly manipulates articulatory space (and timing). Auditory perturbations, on the other side, could exclusively manipulate temporal properties of speech and thereby give insight into their representation and control. With *temporal* auditory feedback perturbations, it is possible to test for a connection between patterns of temporal coordination as defined in the Articulatory/Task-Dynamics framework and the incorporation of auditory feedback mechanisms as incorporated in the DIVA model for their realization.

Delayed auditory feedback studies observed adjustments in speaking rate when the whole auditory feedback is delayed (Yates, 1963). Auditory feedback perturbation now allows researchers to target specific sequences in fluent speech and *focally* alter their temporal dimension.

Only recently have a few studies started to explore the effects of focally applied temporal auditory feedback alterations on speech production. Thereby, the concept of time curtails the possibilities and raises a challenge for performing such shifts: While, in spectral perturbations, shifts in every direction can technically be applied, it is impossible to only compress the duration or even omit segments in real-time. In this case, the signal that should serve as feedback would not have been produced yet. Since time-travel into the future is not (yet) possible, real-time temporal alterations hold the constraint that manipulations can only compress parts of a signal when previous parts have been stretched. By first stretching and then compressing the signal, it is ensured that the signal that should serve as auditory feedback is already produced and that the signal after both manipulation directions matches the real-time of production again. An initial study by Mitsuya *et al.* (2014) bypassed this circumstance by altering the duration of distinctive voice onset time (VOT) in a minimal pair *offline.* They fed back pre-recorded versions of the other token of the minimal pair than the one produced. The speakers adapted for the erroneous category-crossing feedback by changing the VOT in production. Another study by Cai *et al.* (2011) did not explicitly alter the duration of segments but shifted the spectral midpoint of a vowel in two directions, causing the vowel target to be perceived earlier or later. Their speakers showed a general reaction by slowing down following segments after random perturbations. Very recently, the study by Floegel *et al.* (2020) altered the absolute durations of sounds by stretching a vowel or a final fricative. They showed that speakers adjust their productions accordingly to the perturbation by shortening the targeted part, causing compensation for the applied shift.

1.3 Objectives of this Thesis

The study by Floegel *et al.* (2020) indicates that temporal adjustments can, in principle, be achieved when the auditory feedback is altered in real-time, at least in one direction (by shortening segments).

But how are temporal adjustments in reaction to shifted auditory feedback possible? As stated above, the gestural organization described in Articulatory Phonology/Task-Dynamics does not incorporate auditory feedback to control timing mechanisms, and the DIVA model is not explicit about the initial temporal organization. Do speakers adjust for temporal shifts regardless of the coordinative pattern? The temporal coordination of gestures differs dependent on the prosodic structure. For example, syllable onsets show different patterns of overlap and coordination with the syllable's nucleus than syllable codas of the same phonetic structure. Further, the lexical stress pattern causes greater durations of stressed syllables compared to unstressed syllables of the same phonetic structure.

These coordinative patterns are, according to Articulatory Phonology/Task-Dynamics, pre-planned and should not be guided by, or adjustable based on, the auditory feedback.

The subsequent investigations presented in this thesis aim to explore the feedforward stability of structural prosodic timing patterns when the auditory feedback is temporally altered in real-time. The ability of the motor system to adjust bi-directionally will be tested in an adaptation paradigm when segments will be stretched and compressed in real-time. Responses to auditory feedback shifts can be expected to shed light on the contribution of auditory feedback for the control and planning of speech timing. Further, the findings shall pave the way for a comprehensive speech production model that eventually incorporates coordinative timing mechanisms with a concept about the role of auditory feedback and feedforward mechanisms.

In the first two experiments (chapter 2 and chapter 3), two temporal auditory feedback perturbation experiments will be reported that test for the malleability of prosodic timing patterns when their structure is auditorily perturbed. Thereby, specifically targeted segments of a fluent speech sequence will be stretched and compressed in real-time. The third experiment (chapter 4) examines possible connections between responses to temporal auditory feedback perturbation and individual perceptual and motor executive abilities.

1.3.1 *Chapter two: Temporal Perturbation and Syllable Structure*

The first study (presented in chapter 2) addresses syllable structure as an influencing factor for the temporal coordination of speech. The Articulatory Phonology/Task-Dynamics framework suggests that gestural overlap is greater in onset consonants than in coda consonants and that onset and coda consonants are coordinated differently with the syllable's nucleus. Syllable onsets share a single point of coordination with the following vowel independent of how complex the onset might be, while in codas, every consonant is sequentially timed with the preceding one. This results in greater articulatory stability of syllable onsets compared to syllable codas (Browman and Goldstein, 1988; Pouplier, 2012). The first experiment examines the stability of onsets and codas when the auditory feedback is temporally altered. Similar sounds in onset+vowel and vowel+coda position will be perturbed, whereby the first segment will always be stretched and the second segment compressed. This method firstly tests whether speakers generally compensate for an introduced temporal shift as they do for spectral alterations. Secondly, this chapter explores whether the structural stability of onsets leads to different response patterns in the face of the perturbation compared to the manipulation of the coda. Suppose auditory feedback is indeed used to monitor or/and update timing mechanisms in speech, less articulatory adjustments to manipulated onsets than to codas can be expected due to the onset's greater articulatory entrenchment. The findings will be discussed with reference to existing speech production models and findings on timing mechanisms from related areas such as music research. A version of this chapter has been published in the *Journal of the Acoustical Society of America* and figures are reproduced from Oschkinat, M., and Hoole, P. (2020). "Compensation to real-time temporal auditory feedback perturbation depends on syllable position," J. Acoust. Soc. Am. 148, 1478-1495, with permission of the Acoustical Society of America.

1.3.2 Chapter three: Temporal Perturbation and Lexical Stress

The second study, presented in chapter 3, approaches lexical stress as another prosodic phenomenon that strongly affects the temporal organization in fluent speech. In German words with two similar syllables that carry the same nucleus vowel, the stressed nucleus will always be longer than the unstressed one. With a similar design as in chapter 2, the onset+vowel of two phonetically similar syllables in a three-syllabic word will be temporally altered. The word-initial syllable will be unstressed, and the word-medial syllable stressed, followed by a third syllable of different phonetic structure. Analogously to the onset condition in chapter 2, the onsets will be stretched and the vowels compressed. The perturbation of the second syllable weakens the stress pattern with the compression of the vowel. Therefore, we hypothesize that speakers adjust more for the perturbation in the stressed syllable than for the same perturbation in the unstressed syllable to preserve the intended lexical stress pattern. In this chapter, all segments of the target word will be examined which gives insight into global and local patterns of response to perturbed timing.

Both perturbation experiments follow a very similar procedure. However, since a version of the third chapter is published as a journal article in the *Journal of Phonetics* (Oschkinat and Hoole, 2022), the description of the theoretical background and methods section might appear repetitive in this thesis. In comparison with the first study presented in chapter 2, it is worth pointing out that in the methods section of the third chapter, the analyses and results sections are more extensive than the analyses in chapter 2 and partially differ from those in some points (e.g., sections 3.3.1 and 3.3.3). Section 3.3.2 follows the same principles as section 2.4.2.2 in chapter 2. Additionally, apart from the analysis of durational properties, chapter 3 also considers somewhat neglected potential trade-off effects in suprasegmental cues. That is, while in chapter 3 duration as a segmental marker of stress is perturbed, the intensity, and spectral skewness of the perturbed sequences will be analyzed for further insight into the exchangeability of speech properties with one another when one of them is being perturbed (section 3.3.4). In chapter 3, our first study (chapter 2) will be referred to frequently, but in its Journal paper appearance (Oschkinat and Hoole, 2020) to allow for a more integrated writing style. The version of the study presented in chapter 3 differs from the journal article in additional examinations of changes in f0 to be found in Appendix (D).

1.3.3 Chapter four: Temporal Perturbation, Perception and Motor Execution

Chapters 2 and 3 suggest that *structural* feedforward stability plays a role in responses to temporally altered auditory feedback, with syllable onsets being more entrenched in the motor system and less malleable through (erroneous) auditory feedback. Further, chapter 3 assumes temporal representations of stressed vowels to show less tolerance towards shorter productions than their unstressed counterparts. The investigation of structural stability/structurally-induced behavior during temporal auditory feedback perturbation has raised the question of whether *individual* differences in feedforward stability might also cause systematic differences in responses to temporally altered auditory feedback. Individual feedforward stability has not been considered much in previous investigations as an influencing factor for speech production. Individual perceptual abilities, on the other hand, have been of significant interest in connection with spectral auditory feedback perturbation studies. Previous studies established a link between the ability to distinguish subtle changes in auditory stimuli (*auditory acuity*) and the size of speech spectral speech targets. Further, auditory acuity was linked systematically to responses to spectral feedback perturbations: Speakers with a better auditory acuity were found to compensate more for spectral shifts, suggesting that the auditory mismatch is more precisely perceived and counteraction can be more effective (Villacorta *et al.*, 2005; Villacorta *et al.*, 2007; Nault and Munhall, 2020).

The fourth chapter picks up on this point and inquires into *individual* differences in feedforward stability, meaning the ability to produce motor actions as precisely as possible, and further into individual auditory acuity regarding temporal discrimination abilities. The examination of individual feedforward stability has, to our knowledge, not been considered before and should give insight into individual abilities in motor execution and their relevance for speech production (but see Martin *et al.*, 2018). Auditory acuity has previously been shown to affect reactions to spectral auditory feedback perturbations. However, since temporal auditory feedback perturbation targets different mechanisms than spectral perturbations do, it is not self-evident that this relationship naturally also applies to temporal properties of speech.

Especially in chapter 2, predictive timing mechanisms are discussed in connection with related concepts of timing in music production and perception. Neurological approaches have focused much on commonalities and differences between music and speech. In timing, it was shown that deficits in fluent speech production, such as stuttering, go along with deficits in non-speech rhythmic behavior, such as precisely hitting the beat to music (Falk *et al.*, 2015). Both music and speech share the need for a precise interplay between feedforward and (auditory) feedback systems to successfully produce the predicted timing of an event or a sequence of events.

Against this background, in chapter 4, non-speech and speech motor stability and auditory acuity are assessed with a broad set of speech, music, and general perceptual and productional timing tasks. Non-speech motor action will be examined with finger-tapping tasks with and without a pacing event. Perception tasks for measuring auditory acuity will include temporally altered monosyllabic speech stimuli, pure tones, and beat-alignment judgments to speech and music stimuli. The analyses in chapter 4 aim at understanding the contribution of individual differences in motor execution abilities and auditory acuity to reactions to temporally altered auditory feedback. These findings can shed light on the mechanisms involved in building stable temporal representations in speech production and how flexibly patterns of speech timing can be guided via auditory feedback.

To date, very little is known about the reactions to focal temporal auditory feedback perturbation and their underlying mechanisms. Accordingly, even less information is available on which tasks and methods are the most relevant for measuring a relationship between motor execution abilities, perceptual abilities, and reactions to temporal auditory feedback perturbation. Therefore, the experiment in chapter 4 performs a full test battery of motor and perception tasks to get an overview of relevant contributors for testing motor execution abilities and auditory acuity and predict compensatory behavior. Accordingly, it should be stated that the analyses in chapter 4 are exploratory in nature, aiming to provide a lead for future investigations. The data of chapter 4 has been published as a journal article in *Frontiers of Human Neuroscience* (Oschkinat *et al.*, 2022). However, the journal paper version and chapter 4 of this thesis differ significantly in analyses.

1.3.4 Chapter five: Discussion and Outlook

The results of each study will be thoroughly discussed in the respective chapter. In chapter 5 of this thesis, the main results will be summarized. The concluding discussion will focus on commonalities and differences in approaching and evaluating spectral vs. temporal auditory feedback perturbations. Further, it will be discussed what temporal auditory feedback perturbation can teach us about speech timing mechanisms and how the findings open up new perspectives for studying speech (and non-speech) timing mechanisms in healthy and impaired populations.

Chapter 2
Compensation to real-time temporal auditory Feedback Perturbation depends on Syllable Position

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.

2.1 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 timeframe when the auditory feedback of their own speech is altered. In manipulation of 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; 2006b; 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 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 200ms 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 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 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 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 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 e.g. 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 timeframe 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 precise timing of speech

[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.

sounds, particularly in syllable onsets (see e.g. Hubbard, 1998; Max and Gracco, 2005; Etchell *et al.*, 2014), they also show deficits in non-speech 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 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 pre-recorded 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 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 pre-recorded 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 syllable position.

With respect to syllable structure, inter-gestural timing was modeled with *coupled planning oscillators* that may couple mainly in-phase or anti-phase 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 versus coda position. Onsets are coupled anti-phase 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 anti-phase 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 and vowel (CCV) in a consonant-consonant-vowel-consonant (CCVC) syllable (Onset condition), and second to vowel and coda (VCC) in a consonant-vowel-consonant-consonant (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 further 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 bi-directional perturbation applied to multiple segments within a syllable which in turn is part of a complete multisyllabic phrase.

2.2 Methods

2.2.1 *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 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 upwards or downwards 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, while 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 condition can be found in Figure 2.1.A and 2.1.B. 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.

2.2.2 *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).

Since 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 pre-defined 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 to 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 timeframe 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 subjects 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 3cm from the corner of the mouth. The foam eartips ensure that the manipulated feedback rather than airborne sound is predominantly perceived and also minimize the occlusion effect (see Figure 2.1.C for setup). Subjects spoke the target phrase ("besser Pfannkuchen" or "besser Napfkuchen", respectively) 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 2.5s, which allowed the subjects to choose an individual comfortable and natural speaking rate without providing too much time for high variability in 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 24ms. The playback volume was set to a comfortable level but loud enough to ensure that they did not hear their own airborne sound. The 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 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, Figure 2.1.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.

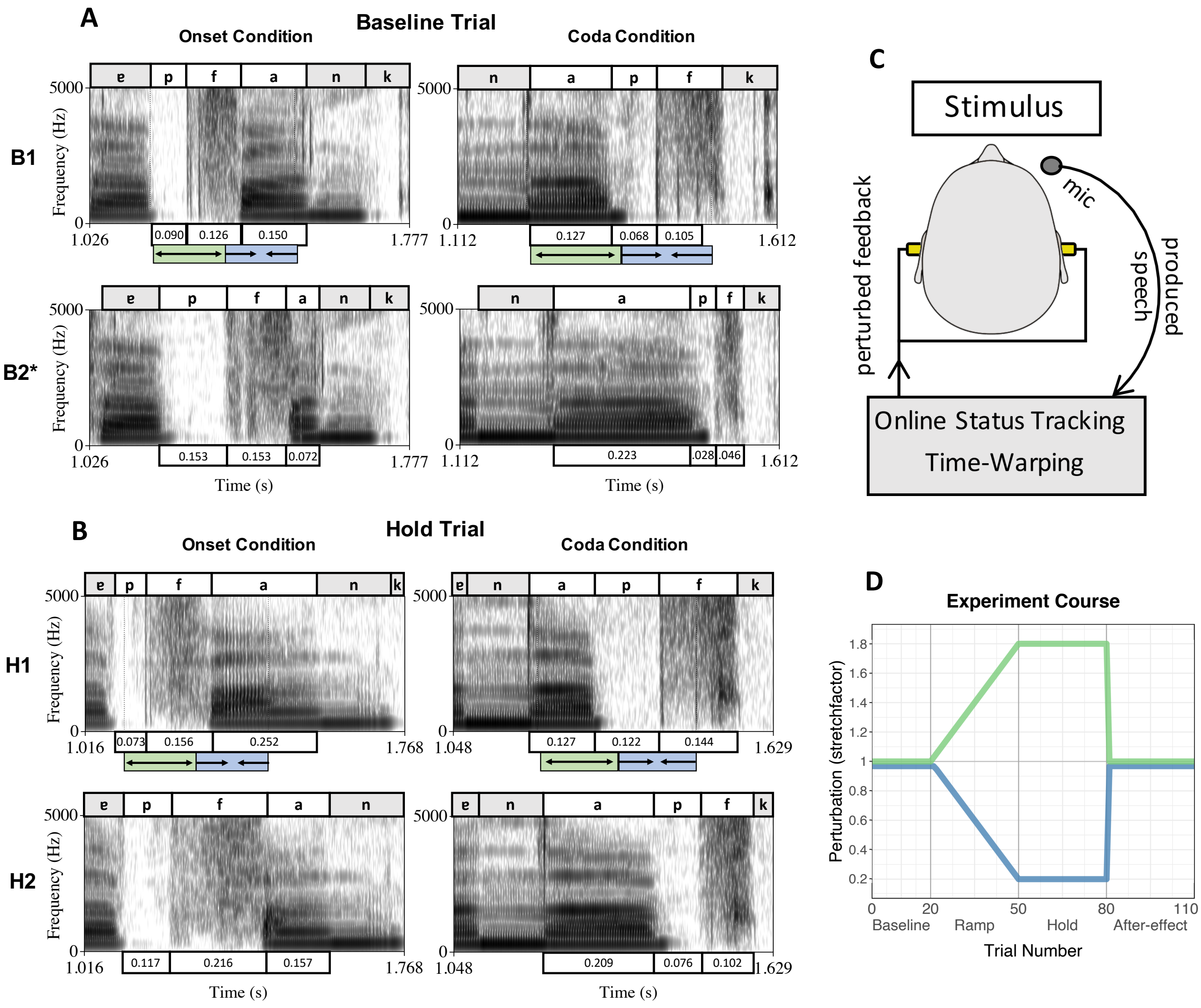


Figure 2.1: **(A)** Spectrograms of a baseline trial per condition of one subject. The onset perturbation condition appears in the left panels ("besser Pfannkuchen," bold section visible in the spectrograms) and the coda perturbation condition appears in the right panels ("besser Napfkuchen," bold section 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 spectrograms. 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 and the coda perturbation condition appears in the right panels, and segments of interest and their durations are marked similar as in (A). Note that productions in the upper panels might already be produced compensatorily. 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).
(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.

2.3 Analyses

2.3.1 *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 was available for 34 subjects for the Onset condition (mean: 23y, 27f) and 33 subjects for the Coda condition (mean: 23y, 27f). Twenty-eight of those subjects provided data for both perturbation conditions. From a total of 3740 trials in the Onset condition (34 subjects x 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 x 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.

2.3.2 *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 (for example Mitsuya *et al.*, 2014 excluded the first 10 trials). In the current study, higher variability in production during the first 9 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 behavior. Figure 2.2 visualizes the produced normalized relative durations averaged over all subjects of the CC segment /pf/ (green dots) and the vowel /a/ (blue rhombuses, color online). The baseline mean (calculated from trial 10 to 20) represents the zero line. Positive values indicate a lengthening, negative values a shortening relative to baseline productions. The spoken signal is shown in solid colors, the perturbed (heard) signal with higher transparency. Please note that the perturbed/heard signal does not represent a one-to-one mapping of the applied perturbation since it is possibly diminished by compensatory behavior. The difference between spoken (solid) and heard (transparent) signal 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.

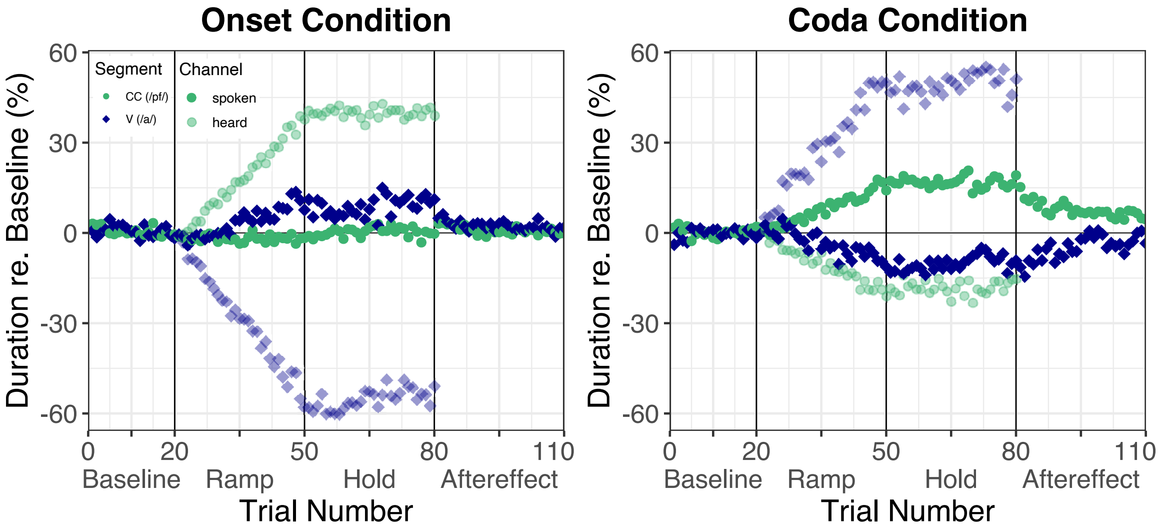


Figure 2.2: Normalized relative durations averaged over all subjects (n=34 for the Onset condition, n=33 for the Coda condition) per trial. The vowel /a/ is shown in blue rhombuses and CC /pf/ is shown in green round dots. The spoken signal is shown in solid colors and the perturbed (heard) signal is shown with higher transparency. The left panel visualizes the Onset condition and the right panel visualizes the Coda condition.

2.4 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 phase.

Firstly, the ramp phase provides information about the reaction threshold and sensitivity to gradually increased perturbation (section 2.4.1). Secondly, 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 (section 2.4.2.1). Additionally, the reaction magnitude of the whole perturbed segment (CCV and VCC) is set in relation to the applied amount of perturbation (section 2.4.2.2). Lastly, the aftereffect phase analysis provides the span of trials for which reactions may persist when normal feedback is abruptly restored (section 2.4.3).

Each phase was modeled individually to capture within-phase behavior. Modeling over phase boundaries (statistically or visually) could distort timepoint specific effects related to the very different perturbation status of 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.

2.4.1 *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 non-linear 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 non-linear relationships in the data by relying on parametric terms and smooth terms. The smooth terms define the shape of the fitted curve by adding up basis functions to a more complex curve until it fits the data properly. Unlike 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 non-linear 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 (trial 21 to 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 trial number (non-linear effect of trial number on normalized relative duration) by segment; 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 (Figure 2.3). 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 2.3 shows the produced differences over the ramp phase and the span of significant deviation from the baseline mean (0). With a significant effect around the same trial for the vowel in onset and Coda condition, the sensitivity to vowel perturbation seems not to be influenced by perturbation direction (stretching or compressing) or whether it is the first or second perturbed segment.

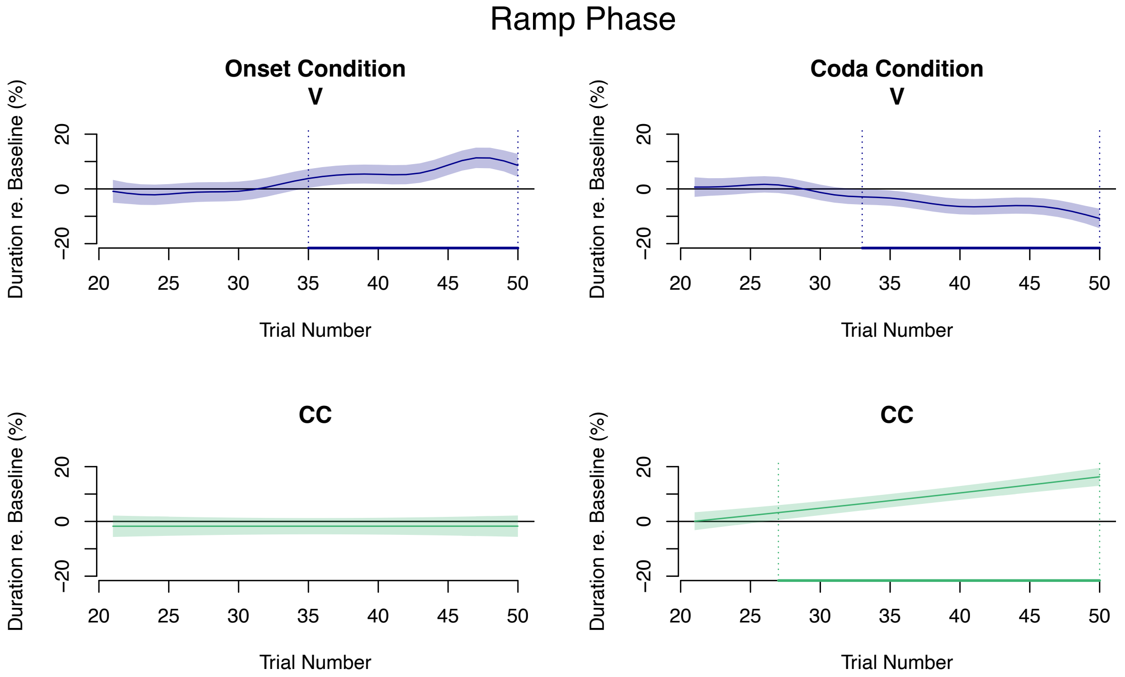


Figure 2.3: GAMM fits of the ramp phase, including random effects and confidence intervals (95%). The Onset condition appears in the left panels (34 subjects) and the Coda condition appears in the right panels (33 subjects). CC fits are shown in green and vowel fits are shown in blue. Dotted vertical lines and thick horizontal lines mark the significance from zero for each sound.

2.4.2 *Hold Phase*

2.4.2.1 *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 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 a by-subject intercept and a random slope for phase and for 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 condition (α = 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 2.4 summarizes the durations in the hold phase relative to the baseline mean (zero).

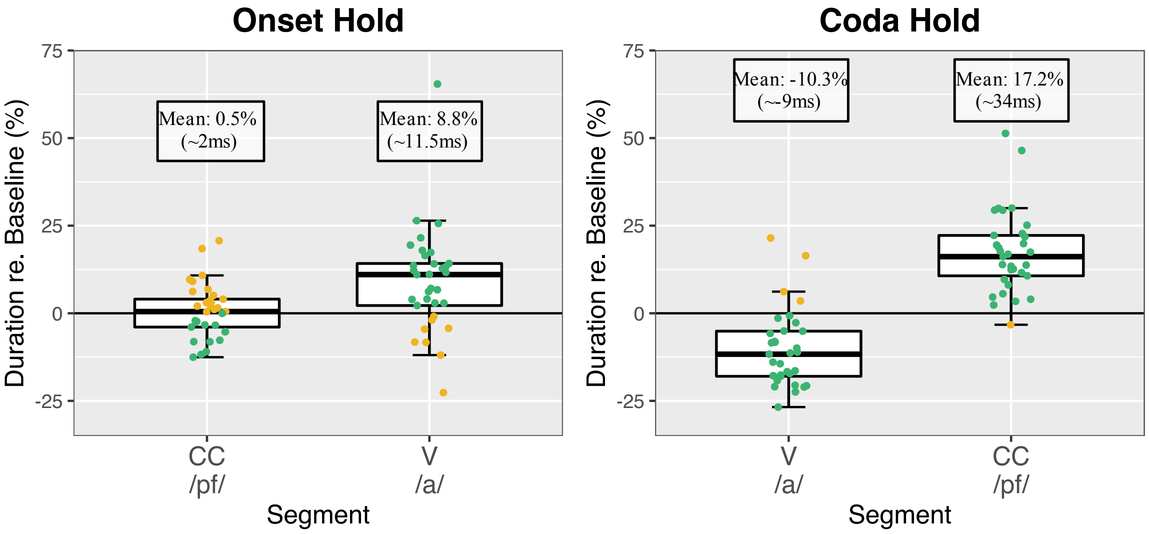


Figure 2.4: Normalized relative durations in the hold phase relative to the baseline mean (0) for vowel /a/ and CC /pf/ in the onset perturbation condition (34 subjects, left panel) and coda perturbation condition (33 subjects, right panel). Boxes correspond to the first and third quartiles and bars 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.

For completeness, linear mixed models with similar specifications as 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 Bonferroni-corrected significance level revealed results for the single sounds. For the onset consonant sequence /pf/ (Onset condition) the model reported a non-significant average shortening of 2.7% (~3ms) for C1 /p/ (estimate = -2.72; SE = 1.73; df = 54.88; *t-ratio* = -1.57), and a non-significant lengthening of C2 /f/ of 3.8% (~5ms; 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% (~15ms) for C1 /p/ (estimate = 18.71; SE = 2.47; df = 43.59; *t-ratio* = 7.58; $p < 0.025$), and 17.4% (~19ms) for C2 /f/ (estimate = 17.45; SE = 2.47; df = 43.58; *t-ratio* = 7.07; $p < 0.025$).

Figure 2.5 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 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 Figure 2.1 for visualization of the fit of the perturbation section). Figure 2.5 indicates that in the Onset condition both single consonants have been stretched in perturbation (transparent dots, squares, and triangles). In productions, C1 has rather been 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 (since both the spoken and the heard signal have approximately the same durations, solid and transparent blue squares), while 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 the discussion (section 2.5).

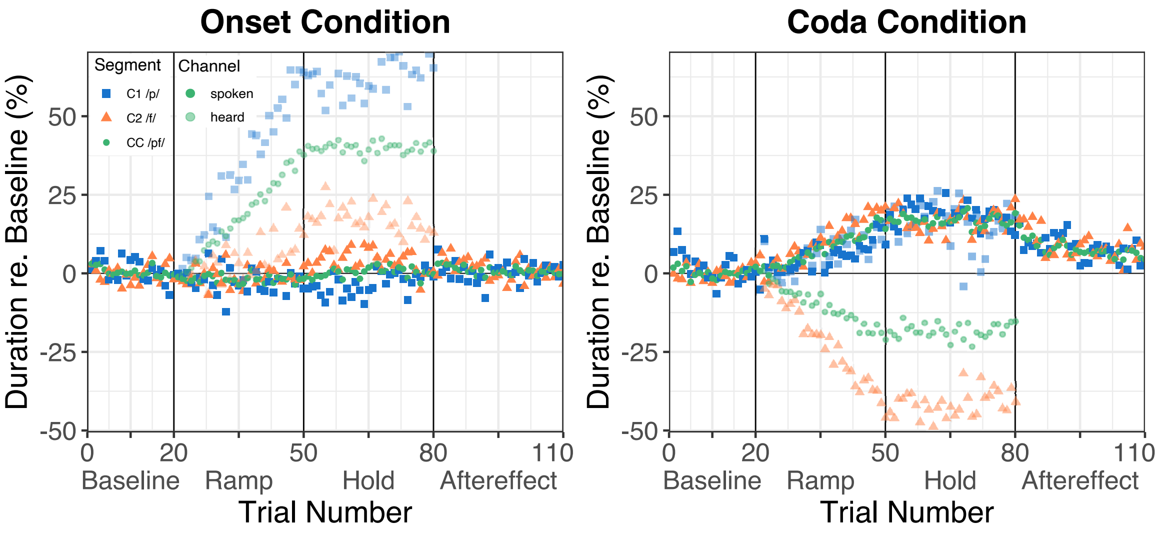


Figure 2.5: Normalized relative durations averaged over all subjects (n = 34 for the Onset condition, n = 33 for the Coda condition) per trial. The CC /pf/ is shown in green round dots, C1 /p/ is shown in blue squares, and C2 /f/ is shown in orange triangles. The spoken signal is shown in solid colors, 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.

2.4.2.2 Compensation relative to perturbation

The analysis of duration differences between 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 reaction relative to perturbation and allow a comparison between onset and coda perturbation for the whole perturbed sequence (CCV/VCC). To ensure a clean comparison between onset and Coda condition, 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 Onset condition and V for Coda condition) are on the x-axis and the durations of the second segment (V for Onset condition and CC for Coda condition) are on the y-axis (for visualization see Figure 2.6.A and 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 9 baseline trials were excluded. Examples of the signals can be found in Figure 2.1: Figure 2.1.A shows the signal of a baseline trial spoken by a subject (B1) and below the simulated perturbation of that signal (B2*). Figure 2.1.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*|, Figure 2.6.A and B, dashed line) and perturbation on a signal that perhaps includes a reaction in the hold phase (Euclidian distance |H1-H2|, Figure 2.6.A and B, dashed line, see equation 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 signal 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 equation 2) forming our *compensation* values.

$$(1) \qquad mean\ perturbation = \frac{|B1-B2|+|H1-H2|}{2}$$

$$(2) \qquad compensation = 1 - \left(\frac{|B1-H2|}{mean\ pert.}\right) * 100$$

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 Figure 2.6.C)

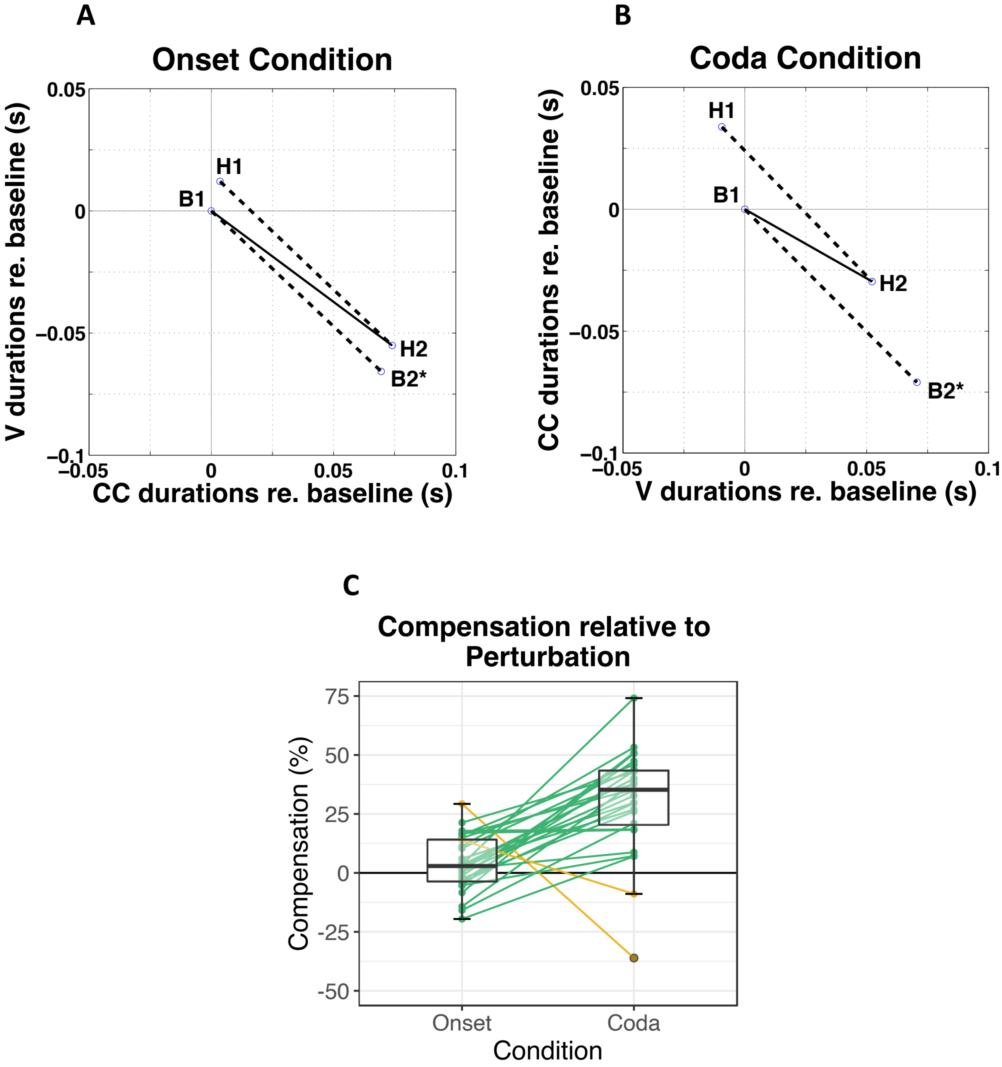


Figure 2.6: **(A)** and **(B)** show mean durations (s) of both segments of interest (V /a/ and CC /pf/) over 28 subjects per perturbation condition relative to the baseline mean (0/0). The first segment of the perturbation section is on the *x*-axis and the second segment of the perturbation section 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* andH2 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 ($n = 2$).

2.4.3 *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 where auditory feedback was abruptly restored, we expected the behavioral data to show either linear or non-linear functions peaking off from maximum compensation towards the baseline mean again. To capture these possible patterns, GAMMS were fitted over all trials of the aftereffect phase (trials 81 to 110).

As previously done for ramp phase examination (see section 2.4.1), 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 (non-linear effect of trial number on normalized relative duration) by segment; 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 the CC segment (which was expected for the CC segment, since 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 Figure 2.7.

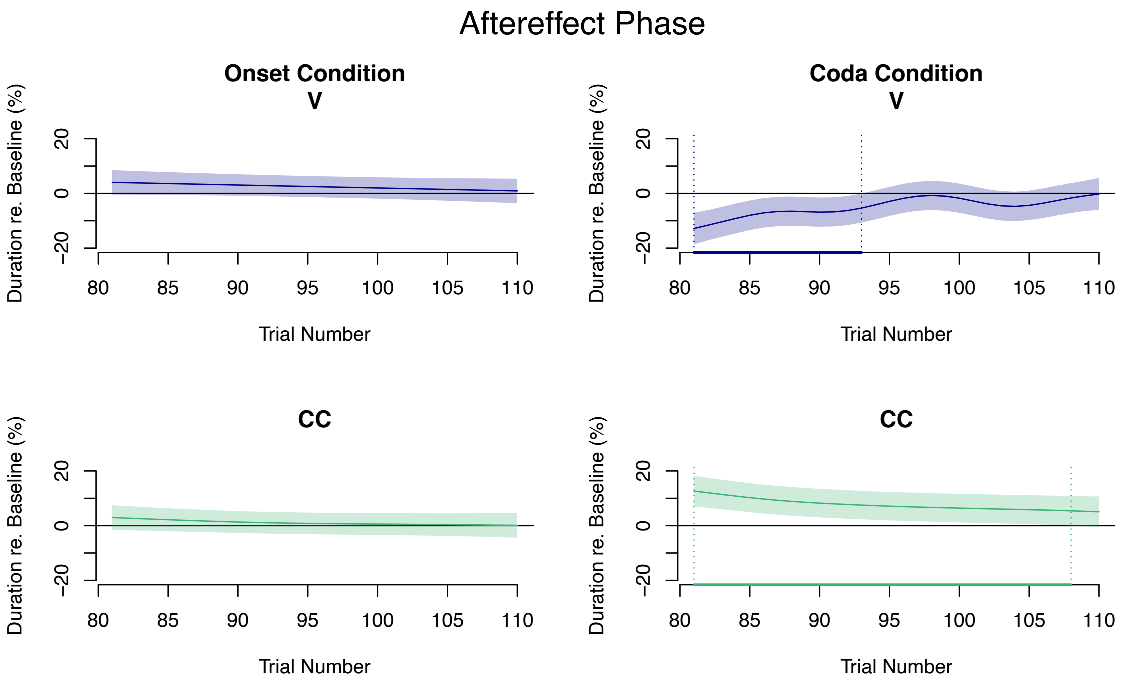


Figure 2.7: GAMM fits of the aftereffect phase, including random effects and confidence intervals (95%). The Onset condition appears in the left panels (34 subjects) and the Coda condition appears in the right panels (33 subjects). The CC fits are shown in green and the vowel fits are shown in blue. Dotted vertical lines and thick horizontal lines mark the significant deviation from zero for each sound.

2.5 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 perturbation direction (stretching or compressing) or whether it is the first segment (Coda condition) or second segment (Onset condition) of the perturbed section.

2.5.1 *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 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 since 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 partly be caused 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 Figure 2.1) that the total duration of the perturbation section was of the order of up to 300ms. Thus, the second half (where perturbatory compression is applied) is about 150ms 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 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 indicates that the nature of the reaction to temporal perturbation is affected by 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 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 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 Mitsuya *et al.* (2014), in the sense that subjects showed a certain amount of compensatory shortening for the initial plosive C1 /p/ (Figure 2.5, blue solid squares). Then again, subjects rather followed the perturbation direction in production by lengthening C2 /f/ (Figure 2.5, orange solid triangles). Taken together, this resulted in an (almost) equal duration of the whole CC onset segment throughout the experiment (Figure 2.5, 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/ (Figure 2.5, 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 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 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. This in short 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 onset and coda position.

2.5.2 *Sensory Interdependence and Feedback Processing*

When comparing onset and coda behavior it remains a concern that they have been treated differently in perturbation. While 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 condition 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 (section 2.4.2.2). The compensation to onset perturbation was overall smaller 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% to 30% (Max *et al.*, 2003; Purcell and Munhall, 2006a; MacDonald *et al.*, 2010; 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 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; 2006a; 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 where 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 prosodic position within the syllable. A greater reliance on somatosensory feedback of onsets could explain their greater resistance to updating of 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 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 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, while 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 1) the suggestion that speakers do not use auditory feedback for the timing of onsets to the same degree as they do for codas and 2) that 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.

2.5.3 *Nature of Timing Mechanisms in Speech and Non-speech*

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 non-speech 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, 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 timeframe. It was found that event-based 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 rather 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 of 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 that 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 non-verbal 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, the latter being influenced to a greater extent by auditory feedback information.

Certainly, these assumptions need further verification e.g., 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.

2.5.4 *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 control and planning level. 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 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.

2.5.5 *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 Figure 2.4 and Figure 2.6.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 rather followed the direction of the perturbation (see e.g. Burnett *et al.*, 1998; Hain *et al.*, 2000 for pitch perturbation; e.g. Purcell and Munhall, 2006b for formant perturbation; e.g. 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 Figure 2.4 (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 perturbation) 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 (section 2.4.2.2). We will not explore this further here, but individual differences in compensatory response to temporal perturbation and their origin could be a specific focus of interest and linked to individual rhythmic and temporal discrimination abilities in future investigations.

Chapter 3 Reactive Feedback Control and Adaptation to perturbed Speech Timing in Stressed and Unstressed Syllables

Abstract

This study examines speakers' reaction to focally applied temporal real-time auditory feedback perturbation in a word-initial unstressed syllable (Unstressed condition) and a similar word-medial stressed syllable (Stressed condition) in a three-syllabic word. Speakers compensate locally in both conditions for the perturbed syllable's nucleus (V; compressed by the perturbation) but not for the complex onsets (CC; stretched by the perturbation). The perturbation of the first, unstressed syllable causes a global slowing down of all segments following the perturbation (syllable two and three), while the perturbation in the Stressed condition elicits local adjustments only in the perturbed (second) syllable. When viewed in a larger prosodic context, the timing strategy in the Unstressed condition indicates that speakers aim to keep relative durations within the word constant when the word-initial onset is auditorily stretched, leading to a compensatory pattern for both CC and V in word-proportional durations. In the Stressed condition, increasing the stressed vowel's duration seems to be of the highest priority, causing all other segments to take up a shorter portion within the word. Adaptation effects of the stressed vowel indicate a durational representation on the segment level. Further adaptation effects additionally suggest a representation of timing/coordination in larger prosodic units. Complementary investigation of aperiodicity, spectral skewness, and intensity (RMS) indicates that spectral properties can change along with compensatorily increased duration.

3.1 Introduction

Speech production requires a precise interplay of feedforward and sensory feedback mechanisms. Perturbations of auditory feedback examine this interplay by manipulating acoustic parameters of a spoken sequence online. In many auditory feedback perturbation studies, speakers produce an isolated vowel, a word, or a phrase while one or more spectral parameters in their auditory feedback are altered in real-time. The initial study by Houde and Jordan (1998; 2002), for example, raised the first formant (F1) frequency in productions of "pep" (/pɛp/), leading to percepts that sounded like "pap" (/pap/) to the speaker. Consequently, speakers started to *compensate* for the received feedback mismatch by lowering F1, leading to productions closer to "pip" (/pɪp/). A manifold body of research has shown that speakers compensate for shifts in the spectral domain. The current study aims at adding to our understanding of the contribution of auditory feedback to *timing mechanisms* in planning and monitoring fluent speech by perturbing speech timing in real-time.

Spectral perturbations have shown that speakers integrate auditory feedback at the control and planning levels, whereby these two levels are typically targeted with different experimental paradigms. In unexpected perturbations of random trials, reactions emerged in the perturbed trial with a latency of ~120-200 ms after perturbation onset, indicating online compensation in online/moment-to-moment control of the ongoing speech sequence (Burnett *et al.*, 1998; Xu *et al.*, 2004; Purcell and Munhall, 2006b; Tourville *et al.*, 2008; Niziolek and Guenther, 2013). However, not every online reaction is compensatory. An online response that is not necessarily compensatory in direction, and might not occur directly at the perturbation site itself, will be referred to as *reactive feedback control*. Consistently applied perturbations over many adjacent trials, on the other hand, can cause speakers to adjust following productions of the perturbed segment. Adjustments in future unperturbed productions indicate an update of motor representations at the planning level *(adaptation)* (Houde and Jordan, 1998; Purcell and Munhall, 2006a; Mitsuya *et al.*, 2011).

In alterations of formant frequencies, shifts mainly targeted isolated vowels or vowels embedded in monosyllabic real-words (Houde and Jordan, 1998; 2002; Villacorta *et al.*, 2007; Mitsuya *et al.*, 2011). Consequently, perturbations of vowels in monosyllabic words give insight into the nucleus' control and representation in stressed syllables. Only a few studies also perturbed the center of gravity (COG) of fricatives in monosyllabic words with the finding that speakers also compensate in onset (Shiller *et al.*, 2009) and coda (Klein *et al.*, 2019) position. Beyond that, very little is known about how prosodic structures such as syllable-, word-, or phrase-complexity shape the control and representation of sounds in higher prosodic units. One study by Lametti *et al.* (2018) examined sensorimotor learning during formant perturbations in entire sentences. They found adaptation in the context of the perturbed sentence and transferred adaptation in future productions of single words, indicating a shared representation for vowels in single-word representation and higher prosodic organization.

The recent study by Bakst and Niziolek (2021) brought prosodic factors more into focus by investigating responses to shifted F1 in words with different stress patterns. Their paradigm not only studied the interplay of stress pattern and syllable position but also explored the target specification of schwa. Characteristically, schwa is highly variable in its spectral shape cross-linguistically (e.g., for English: Fowler, 1981a; for Dutch: Koopmans-van Beinum, 1994), mainly due to coarticulation. For this reason, schwa's phonetic representation may be rather unspecified and its realization highly assimilatory. Bakst and Niziolek (2021) increased and decreased F1 in di-syllabic words to test whether schwa has a specified target and whether compensation and adaptation emerge in stressed and unstressed syllables in the first or second position of the word. Their subjects compensated and adapted for the applied shifts in stressed and unstressed syllables, including unstressed syllables with schwa. However, reactions suggested a complex interplay between shift direction, syllable position, and stress pattern.

Besides the studies by Lametti *et al.* (2018) and Bakst and Niziolek (2021), prosodic factors such as stress, accent, and syllable position have not been investigated much in spectral auditory feedback alterations and were therefore not considered as potentially shaping the control or representation of spectral properties of speech. However, prosodic structures are considered highly influential for shaping the control and representation of other aspects of natural speech, such as speech timing and suprasegmental cues.

Therefore, prosodic structures such as stress pattern experienced more attention in manipulations of suprasegmental properties of speech. The study by Natke and Kalveram (2001), for example, shifted the fundamental frequency (f0) of an entire multi-syllabic non-word down in random trials testing for an effect of lexical stress pattern. Their subjects uttered the non-word /tatatas/ either with stress on the first syllable (/'ta:tatas/) or with stress on the second syllable (/ta'ta:tas/). Subjects responded to the shift in the first syllable only when it was long and stressed but not when it was short and unstressed. In the second syllable, effects were significant independently of whether it was long and stressed or short and unstressed. However, the results do not support a straightforward conclusion about compensation in stressed vs. unstressed syllables: With a general reaction latency to unexpected perturbations typically between ~120 and 200 ms after perturbation onset, real-time responses to the shifted f0 should not be expected in short syllables with a mean vowel duration of 125 ms (as reported in Natke and Kalveram, 2001) following an unvoiced plosive.

Another set of studies by Patel and collaborators investigated the exchangeability of emphatic stress cues when one of them is altered. They shifted f0 in a stressed syllable up or down (Patel *et al.*, 2011) or manipulated the intensity of a stressed syllable bidirectionally (Patel *et al.*, 2015) and found increased intensity along with compensation with f0 in their first study, but purely compensation with intensity to perturbed intensity in their later study. These studies indicate that speakers adjust prosodic properties of speech in the face of a perturbation and that some of these parameters interdepend, albeit not straightforwardly.

Stress and syllable position seem to affect reactions to spectral alterations in a complex way. But how about cues that are both segmental and suprasegmental, such as duration? How does stress pattern impact timing mechanisms in speech when the auditory feedback is temporally altered? Prosodic structures such as syllable position, stress or accent, and prosodic boundaries strongly influence temporal properties of sounds and their gestural coordination (Byrd, 1996; Browman and Goldstein, 2000; Byrd and Saltzman, 2003; Cho and Keating, 2009; Goldstein *et al.*, 2009; Nam *et al.*, 2009; Bombien *et al.*, 2010; Byrd and Choi, 2010; Bombien *et al.*, 2013; Goldstein and Pouplier, 2014).

Recent research has shown that when temporal properties of speech, e.g., sound duration, are altered, speakers compensate and adapt much as they adapt for spectral shifts. The study by Mitsuya *et al.* (2014), for example, altered the voice onset time of the word-initial plosive in a word of the minimal pair “dipper/tipper” by feeding back prerecorded tokens of the other word. They found their subjects to compensate and adapt for VOT, although the manipulation did not target the signal online. Floegel *et al.* (2020) stretched final consonants in a word in real-time and observed compensatory shortening while testing the contribution of both cerebral hemispheres for the processing of temporal vs. spectral auditory information. In our previous temporal real-time perturbation study (Oschkinat and Hoole, 2020), we showed that reactions to temporal real-time auditory feedback perturbation depend on position-in-syllable. The data showed compensation and adaptation to perturbed nucleus and perturbed coda durations in a syllable, but no compensation to the perturbed onset in utterance-embedded real-words. We concluded syllable structure to be an influencing factor, with onsets being temporally less malleable due to their assumed greater articulatory stability (Byrd, 1996; Browman and Goldstein, 2000; Goldstein and Pouplier, 2014). The results further suggested that auditory feedback might be used to a greater extent for monitoring and controlling timing of the nucleus and coda than of onsets, since the temporal extent for appropriate syllable timing can be estimated from the already perceived onset duration. These findings were recently endorsed by Karlin *et al.* (2021) who stretched the onset consonants in "zapper, "sapper", and "gapper" and compressed the following vowel. Their speakers did not change the durations of the onset consonants, but compensated and adapted for the following vowel (and adjusted the following consonant /p/). However, by examining the initial consonant duration as a proportion of the perturbed syllable, response patterns indicated opposing reactions to both the initial consonant and the vowel (Karlin *et al.*, 2021), leading to the conclusion that speech timing might rather control for temporal relationships of segments within a higher prosodic unit than absolute durations (Oschkinat and Hoole, 2020; Karlin *et al.*, 2021).

While prosodic effects such as syllable structure might not be a primary subject of interest in spectral feedback alterations, they clearly cannot be disregarded when examining the temporal organization of fluent speech. The findings of Oschkinat and Hoole (2020) and Karlin *et al.* (2021) added substantially to the scarce body of research on the contribution of auditory feedback to the temporal planning and control of fluent speech. To better understand the influence of prosodic factors on the online control and representation of speech timing, the current study examines the role of lexical stress on the temporal organization of fluent speech when speech timing is perturbed. With the current study, we expect focally applied temporal auditory feedback perturbation to shed light on the stability of prosodically determined timing relations and on the extent to which they diverge when speakers compensate.

Syllable structure affects the temporal coordination of gestures on the syllable level. Word stress, in contrast, is lexically anchored and rather affects durations of sounds on the word level. In an unstressed/stressed contrast, stressed syllables are longer than unstressed syllables in many languages (in German: Jessen, 1993; Jessen *et al.*, 1995; in Dutch: Sluijter and van Heuven, 1996; Sluijter *et al.*, 1997; in English: Kochanski *et al.*, 2005; e.g., in Catalan: Astruc and Prieto, 2006; in Austrian German: El Zarka *et al.*, 2015). In German, vowels in unstressed syllables are only phonetically reduced but do not experience phonological neutralization, as seen in other languages such as English (Mooshammer and Geng, 2008). This certainly highlights duration as the most prominent marker for stress to distinguish vowels of the same category in a direct stressed/unstressed comparison context. The stressed syllable of a word can moreover carry an accent in larger prosodic contexts. Accordingly, stress and accent are terms that have been used to distinguish two realizations of emphasis anchored on different prosodic levels. A large body of research has examined the most prominent attributes of stress and accent in production and for perception.

In many cases word stress not only affects duration but also spectral properties of sounds. Along with duration, an increase in overall intensity marks stress as a perceptual cue (Fry, 1955; 1958), with increased intensity in the higher harmonics of stressed syllables (Sluijter and van Heuven, 1996; Sluijter *et al.*, 1997). This effect, however, might not be uniquely attributable to word stress, but might also be found in accented sequences or, more generally speaking, in emphasized sequences due to general more substantial vocal effort (see, e.g., Campbell and Beckman, 1997 for the interplay of accent and stress in English; and El Zarka *et al.*, 2015 for Austrian German). Perception experiments suggested that syllables with higher pitch are more likely to be perceived as stressed (independent of the magnitude of the pitch difference, see e.g.Fry, 1958). Later studies considered pitch markings as a correlate of accent rather than an effect of word stress (e.g., Beckman and Edwards, 1994; Sluijter and van Heuven, 1996; Sluijter *et al.*, 1997) or of general prominence (El Zarka *et al.*, 2015). Kochanski *et al.* (2005) did not find f0 a reliable marker of prominence in production (unlike duration and loudness) and drew the conclusion that speakers (of British English) do not necessarily use pitch to mark prominence in a signal.

Some cues interdepend; for example, duration and loudness are assumed to be processed as a unit but with a dominance of duration over loudness (Turk and Sawusch, 1996). Most studies on stress perception evaluated the perceptual cues of stress by manipulating one or more speech signal parameters offline and presenting them to naive listeners. Accordingly, the speaker and the listener were mostly two different persons, and the presentation of prerecorded tokens decoupled production and perception temporally and intentionally. With the perturbation paradigm of the current study, the speaker is also the listener, and the signal is manipulated in real-time. This approach factors out some aspects that influence the production of prominence, such as predictability (Turk and Shattuck-Hufnagel, 2014) and investigates cues of stress in a barely investigated processing situation. Saying this, the modality and time course of the response is different than in previous studies: the online monitoring of stress in self-generated speech might require other mechanisms than explicit judgments. Reactions to the manipulation are expected to indicate which cues speakers primarily use to implement stress when decoding information plays a minor role.

In the current study, we manipulate CCV syllables with almost identical make-up (/tʃe/) in two different prosodic contexts but similar phonological contexts. Currently, there is still very little known about the reaction patterns of different sounds to focal real-time temporal auditory feedback perturbation. However, our previous investigation (Oschkinat and Hoole, 2020) showed that syllable structure as a prosodic condition shapes the responses. For this reason, the segments and their position within the syllable as well as the lexical item were kept constant in the current study by choosing one word that provides one stressed and one unstressed syllable with similar sounds in both syllables. Both syllables belong to the same German word "Tschetschenen" (/tʃe'tʃeːnən/, Chechens) spoken after the carrier word "besser" (bɛsɐ, better). In "Tschetschenen", the first syllable is unstressed, and the second syllable is stressed. The stressed syllable will also always be the accented syllable due to the fixed target sentence, strictly speaking confounding stress and accent as done in previous studies (see e.g., Bombien *et al.*, 2010). However, the results will be discussed primarily with respect to the word's stress pattern and secondarily will be interpreted with respect to the presence of a nuclear accent on the stressed syllable. Unlike previous perturbation studies that considered stress pattern as influential for responses (e.g., Natke and Kalveram, 2001), alterations will not be globally applied to the utterance but locally to the segments of interest.

The CC onset segment /tʃ/ will be stretched and the following vowel /e/ compressed with real-time auditory feedback manipulation in either the stressed or the unstressed syllable. Similarly to the majority of responses to spectral shifts and recent findings of temporal real-time alterations (Floegel *et al.*, 2020; Oschkinat and Hoole, 2020; Karlin *et al.*, 2021), we assume speakers compensate for the compression of the vowel in the auditory feedback by lengthening the perturbed vowel in production in both perturbation conditions. Since the compression of the vowel in the stressed syllable weakens the lexical stress pattern, we expect articulatory adjustments of a greater extent to the perturbation of the stressed syllable than to the perturbation of the unstressed syllable to maintain the realization of the desired word stress.

Based on the findings of our previous study (Oschkinat and Hoole, 2020), we do not expect significant temporal adjustments to the stretched onset as a whole unit in either the stressed or unstressed syllable but do not rule out possible temporal adjustments of the single consonants C1 and C2. Although /tʃ/ is frequently discussed as a phonemic unit (affricate) rather than a combination of two single phonemes (cluster, see Wiese, 2000, pp. 13-15 for discussion), our previous research has shown that in an onset with more than one consonant both single consonants can show tendencies of different temporal adjustments under perturbed auditory feedback (Oschkinat and Hoole, 2020). Therefore, /tʃ/ will be analyzed on the one hand as one segment, but also, with regard to its phonetic realization, divided into its single components. In fact, the response pattern to perturbation of onset timing can potentially contribute to the discussion on whether /tʃ/ should be treated as mono-phonemic or as two different phonemes.

To date, very little is known about the prosodic level at which temporal properties of speech are stored and planned. For example, the Articulatory Phonology/Task-Dynamics framework provides a plan for temporal coordination of gestures determined by prosodic aspects of fluent speech. Still, it remains unclear to what degree temporal information unfolds only in the coproduction of gestures, or whether single segments of speech such as sounds carry a temporal representation.

Our previous study (Oschkinat and Hoole, 2020) suggested that speech timing is moreover monitored and potentially updated via auditory feedback. The contribution of auditory feedback for timing mechanisms is not elaborated in the Task-Dynamics framework (see Turk and Shattuck-Hufnagel, 2014, for discussion) but was considered essential for planning and controlling (spectral) speech output in the Directions-into-velocities-of-articulators (DIVA) model.

To gain insight into the representation of temporal properties and the contribution of auditory feedback for their control, the analyses will look into absolute sound/segment durations (in ms) (section 3.3.1), sound/segment durations on the syllable level relative to the applied perturbation (section 3.3.2), and sound/segment durations on the word level (normalized by word duration) (section 3.3.3). The investigation on the syllable level will comprise the whole perturbed sequence and allows for a conclusion about the reaction relative to the amount of perturbation. Thereby, a direct comparison between the perturbed stressed and the perturbed unstressed syllable is possible. The analyses of reaction patterns on sound, syllable, and word levels can be expected to shed light on the representation of duration on the sound level or as the result of higher unit prosodic temporal organization (fluent speech). In so doing, this study can contribute to the current discussion on which aspects are essential for comprehensively modeling speech production.

Along with adjustments in temporal control it is possible that other spectral parameters of the signal change as well. Production changes in non-temporal parameters during the temporal perturbation could either be indicative of physiological or psychoacoustical interdependence of one parameter with another (e.g., loudness changes along with changes in duration), or they could counteract the durational perturbation instead of temporal adjustments indicating a trade-off of cues. For present purposes, the intensity of the signal for the nucleus and the fricative of the perturbed syllable will be examined. Further, as a measure of change in the general spectral distribution, we observe the spectral skewness of the vowel and the fricative in the perturbed syllable. For the vowel, aperiodicity will additionally be examined (section 3.3.4.1). Additional analyses of f0 were considered for this study. Such analyses, however, should be sensitive to the intonation pattern speakers produced. While in our study most of the speakers produced a downstepped H* tone on the stressed syllable (a falling intonation pattern), a rising pattern was observed in some speakers or some trials of speakers who mostly produced a falling pattern. Since the non-temporal parameters are potentially relevant to our understanding of interdependencies between stress cues but nonetheless should not distract from the key durational analyses, additional analyses on changes in f0 can be found in Appendix D. Moreover, we do not have a straightforward hypothesis about how f0 would change in production and further cannot neatly attribute changes in f0 to lexical stress.

For the examined parameters intensity, skewness, and aperiodicity, we assume that production differences would comprise greater intensity and less aperiodicity in the vowel as a result of greater emphasis on a vowel that is compressed in the auditory feedback. Further, we assume that a more emphasized vowel is related to greater vocal effort which leads to a more strongly asymmetrical glottal pulse with a shortened closing phase. This, in turn, would generate a less positive skewness (greater intensity in higher frequencies) in the perturbed vowel (Sluijter and van Heuven, 1996). We have no direct hypothesis for the changes in intensity or skewness of the perturbed fricative, as we have no clear hypothesis of how the fricative might behave in temporal terms. It still might be the case that skewness and intensity change along with or arise instead of duration changes as a direct reaction to the applied perturbation. Alternatively, skewness and intensity could be affected by the realization of the following vowel. For this instance, the fricative will additionally be inspected as an exploratory investigation.

The data of the current study reveal speakers' sensitivity to temporal perturbation of a stressed and an unstressed syllable and the influence of auditory feedback on realizing prosodically determined timing. The examination of duration of different prosodic units is expected to give insight into the units of control and the representation of duration as sound specific or as a result of higher prosodic unit organization. This approach, moreover, allows for drawing conclusions about whether similar stressed and unstressed syllables share the same strategies in realizing the intended timing. While duration as the perturbed parameter is the focus of interest, the additional analyses of other spectral parameters give insight into the interdependence and flexibility of different stress markers in production.

3.2 Methods

3.2.1 *Subjects and Setup*

Forty-five monolingual German-speaking adults from the Munich area participated in two experimental conditions. None of them claimed to have any speech or hearing disorders, and all of them were between 18 and 29 years of age (mean age 23.5 y). For the procedure, the experimenter provided the subject with E-A-RTone™ 3A in-ear earphones with E-A-RLINK foam ear tips for perturbed auditory feedback and a Sennheiser H74 headset microphone placed 3 cm from the corner of the mouth. The E-A-RLINK foam ear tips are compressed prior to testing and inserted into the ear canal where they decompress and fill the ear canal. Thereby, they ensure that the manipulated feedback rather than airborne sound is predominantly perceived and minimize the increase at low frequencies of bone-conducted sound that occurs when the ear canal is blocked (occlusion effect, see e.g. Carillo *et al.*, 2020). The experiment was conducted in MATLAB (The MathWorks Inc., 2012a) using the Audapter software package of Cai *et al.* (2008). Initially developed for formant manipulations in utterances with continuous voicing, more recent versions allow for delay shifts, time warping, and pitch shifts in fluent speech (Cai *et al.*, 2011; Tourville *et al.*, 2013). With a maximum delay of unnoticeable 25 ms between spoken signal and received (perturbed) feedback, speakers are mostly unaware that the acoustics of their auditory feedback were manipulated. Subjects received financial compensation for their participation.

3.2.2 *Procedure*

In both perturbation conditions, subjects produced the German word "Tschetschenen" (/tʃeˈtʃeːnən/, Chechens) after the carrier word "besser" (/bɛsɐ/, better). The phrase was lexically presented in a box on a screen. The frame of the box turned green when the recording started and red after 3 seconds signaling the end of a trial. In the first experiment, perturbation targeted the first unstressed syllable (/tʃe/) (Unstressed condition), while in the second experiment, the perturbation targeted the second stressed syllable (/ˈtʃeː/) (Stressed condition). In both perturbation conditions, the Onset CC (/tʃ/)

of the targeted syllable was stretched and the following vowel (/e/) compressed in manipulation. The second syllable vowel /e:/ is longer than the vowel of the first syllable /e/ due to the stress pattern. However, the unstressed vowel is not expected to reduce massively towards another vowel quality, unlike the situation in other languages, such as English (see Appendix A for an overview of produced formants). In each condition, subjects were instructed to speak the phrase "besser Tschetschenen" 110 times resulting in 110 trials per experiment. Half of the subjects started with the Unstressed condition; the other half started with the Stressed condition. Prior to the experiment, speakers were instructed to keep their speech rate as constant as possible throughout the experiment.

The first 20 trials of the experiment served as a baseline and provided authentic feedback. In 30 subsequent trials, the perturbation increased gradually to maximum perturbation (ramp phase), followed by another 30 trials with maximum perturbation (hold phase). For the last 30 trials, regular feedback was restored, allowing for examining learning effects due to the previously experienced persistent feedback alterations (aftereffect phase).

While there is a vast body of research on delayed auditory feedback, there are until today just a few studies that focally altered auditory feedback in the temporal domain. Targeting specific sounds in real-time with temporal manipulation faces more significant challenges than spectral manipulations do since the target of manipulation and its duration change when speakers adjust their productions. One of the challenges is the need to stretch and compress the signal by the same amount. More precisely, if a section of the signal is only stretched, then the part after this section would be overall delayed by the amount of stretching. Exclusively compression, or compression before stretching is technically not possible, because in this case the signal that should serve as feedback after compression has not been produced yet. With stretching and compressing the signal (in this order) each by the same amount, the compression serves as a reversion of the signal to real-time after stretching.

In our implementation, the perturbation always targets the whole syllable by stretching the first part and compressing the second part. In the hold phase with maximum perturbation, perturbation stretched the first half to 1.8 times the input duration and compressed the second half to 0.2 times the input duration, which leads to a constant duration of the whole perturbation section (for visualization see Figure 3.1). Specifically, the present experiment used the time-warping functionality of Audapter, which is based in turn on a phase-vocoder approach. Each input frame is Fourier-transformed into the spectral domain. The frequency and phase representation is interpolated appropriately such that after inverse Fourier transformation back to the time domain the resulting time signal has the desired amount of stretching or compression (see Tourville *et al.*, 2013 for details).

The main focus of manipulation was to target the vowel appropriately with the perturbation. Therefore, the second half of the perturbation section comprised the vowel of the syllable of interest (syllable one or syllable two) to ensure compression. Accordingly, the first half covered the preceding C1 and C2 segments which were stretched. Spectrograms of the manipulation in both conditions are provided in Figure 3.1. Depending on vowel duration, however, C1 and C2 were not always entirely covered by the first half of the perturbation section. In the Unstressed condition the vowel was shorter and therefore more difficult to target precisely in perturbation. In some cases, the following CC segment of the second syllable was partially covered by the perturbation section, thus experiencing some shortening (see Figure 3.1, upper panels).

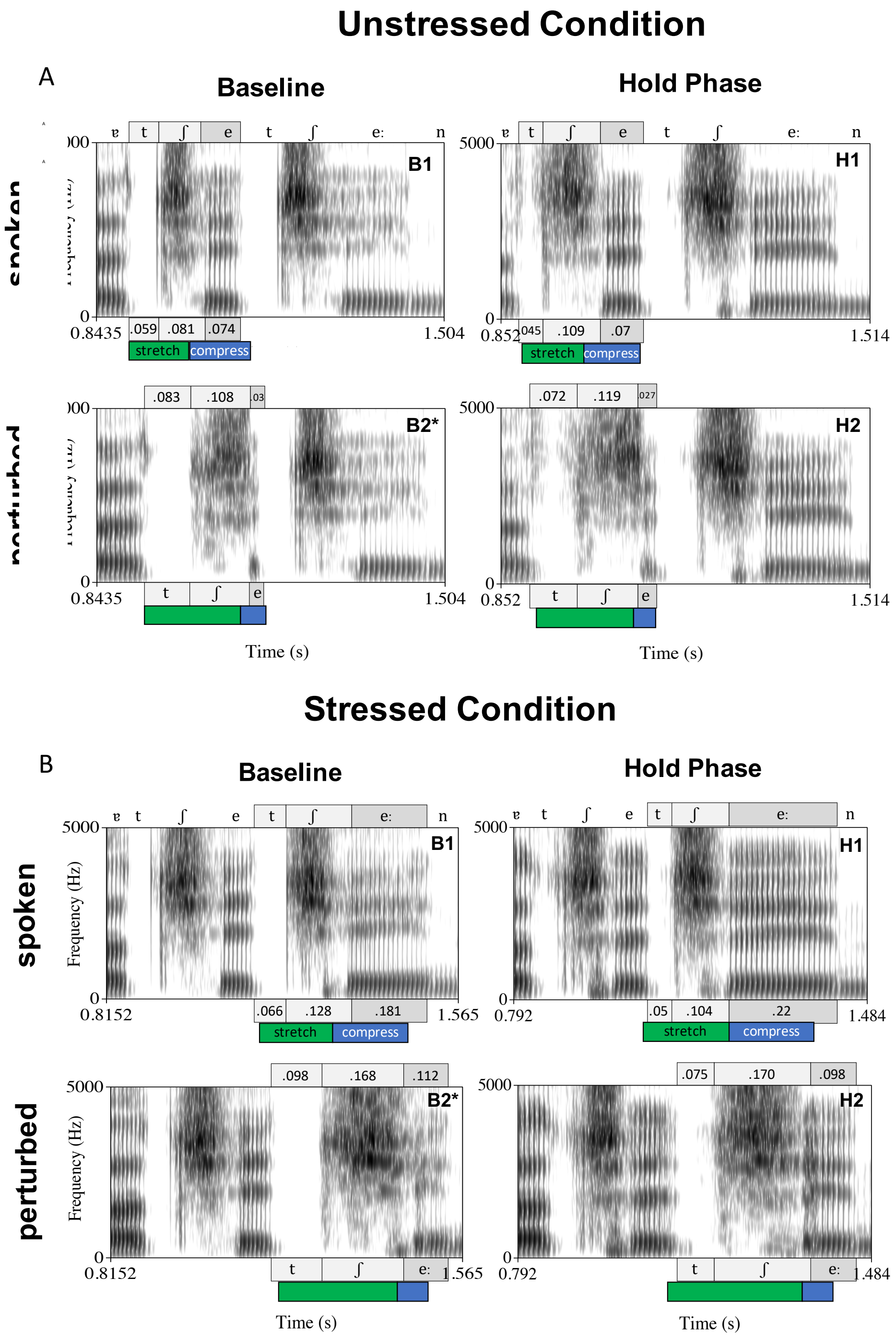


Figure 3.1: Spectrograms of a baseline and a hold phase trial of a male subject for each condition. **(A)** The Unstressed condition (perturbation of syllable 1), **(B)** Stressed condition (perturbation of syllable 2). The upper panels per plot show the spoken signal of one baseline trial (**B1**) a and Hold phase trial **(H1),** and the lower panels show a (**simulated)* maximum perturbation of the same trial in the baseline (**B2***) and Hold phase **(H2),** respectively. 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 and their durations shown above/below the spectrograms. The perturbed segments are marked in grey ([t] and [ʃ] in lighter grey, the vowel in darker grey). Below the targeted segments the perturbation section for the respective trial is shown. The green part marks the first half of the perturbation section covering the signal that is stretched in perturbation, the blue part marks the second half of the perturbation section that is compressed in perturbation. Note that the perturbed signal (B2*) includes the Audapter delay of 24 ms.

3.2.3 *Pretest and Online Status Tracking (OST)*

Before the actual testing session of a perturbation condition started, the subject underwent a pretest per perturbation condition. The pretest consisted of 10 to 20 tokens of the baseline condition (no perturbation), depending on how fast the subject established a consistent speech style and felt comfortable. Speakers were instructed to speak naturally but as constantly as possible without any intended variation in speaking style. This pretest served to get the subject used to the procedure and subsequently measure the mean vowel duration of the last 10 stable productions. Twice the mean vowel duration served as an individual duration of the perturbation section. The second half of that section covered the vowel and the first half the preceding signal.

To target the part of the signal that should be altered, Audapter comes with an online status tracking (OST), which evaluates the status of the spoken signal based on predetermined thresholds for the RMS or the pre-emphasized RMS of the amplitude. Thresholds have to be determined according to the spoken sequence. For example, vowels can be detected by defining high thresholds in the RMS of the amplitude, fricatives can be detected by determining thresholds of the pre-emphasized RMS curve of the signal. For the purposes of the current study, the carrier word "besser" was chosen as it provides vowels and fricatives that are well detectable by Audapter's online status tracking. For the manipulation of the first syllable (/tʃe/), the OST thresholds were adjusted to fit the word "besser" (/bɛsɐ/), with the onset of the second vowel in "besser" (/ɐ/) as the last detected OST state. For each speaker, an individual duration (elapsed time) was implemented measured from this last detected OST state to the start of the closure in [t]. For targeting the second syllable (/ˈtʃeː/), the automated OST triggered until the onset of the vowel /e/ in the first syllable of "Tschetschenen" (/tʃe/), and from that point to the start of the closure of the second [t] an individual duration (elapsed time) was measured. The experimenter implemented the individual perturbation section's duration and the elapsed time duration into each subject's test procedure per perturbation condition before the test started.

3.2.4 Data Exclusion

For precise perturbation of the intended sequences, well-functioning OST-tracking is crucial, as well as the implementation of the *elapsed time* duration and the duration of the *perturbation section* in our paradigm. However, this implementation did not lead to the intended perturbation when subjects changed their productions in some unexpected way or showed very high variability between trials. For those reasons, some subjects had to be excluded from further calculations.

One reason for exclusion, especially in perturbation of the first syllable, was the insertion of a pause between the two words of the utterance, which resulted in a poor fit of the perturbation section or even caused the whole perturbation section to lie within that pause (which could indicate an avoiding strategy). Further, some subjects strongly lengthened the onset CC in production, which caused the /e/ to lie outside the perturbation section. The latter points to one special case we do not capture with the data of the current study: Extensive lengthening of the CC segment in production causes the vowel (especially in the Unstressed condition) to lie outside the area of perturbation, which leads to the exclusion of those subjects. However, only two subjects strongly lengthened CC (or one of the two consonants) in a way that led to exclusion. An example of a bad fit of the perturbation section because of intensive onset lengthening in production can be found in Appendix B.

An automated Matlab script identified and removed trials in the ramp and hold phase where the vowel did not lie within the second half of the perturbation section for each of the perturbation conditions. If a subject had less than 16 acceptable trials in both the ramp and hold phase, the whole subject was removed from calculations of that condition.

One other subject was excluded because of a very slow and unnatural speaking style in both perturbation conditions. Another subject was removed due to the incorrect realization of the stress pattern (stress on the first syllable). Two more subjects had to be excluded as they probably showed perturbation-related reactions that were, however, not evaluable as such with the following statistical methods. One of them started to stress the first (unstressed) syllable during the Unstressed condition in the hold phase and continued with that stress pattern for the rest of the experiment, including the second (Stressed) perturbation condition. Another subject started to show stuttering-like symptoms by frequently repeating the third syllable in perturbed trials

("Tschetschenenen"). In total, 14 subjects qualified themselves as outliers based on the reasons stated above in the Unstressed condition (syllable 1), and four subjects in the Stressed condition (syllable 2). Since this resulted in a very unbalanced dataset of subjects between the perturbation conditions, we decided to include only subjects with data in both perturbation conditions into all following calculations, resulting in 30 subjects per perturbation condition.

3.3 Analyses and Results

All segment durations of the target word "Tschetschenen" were hand-segmented by research assistants (naïve to the purpose of the experiment) in praat. The following analyses will be performed on parameters extracted from these segment-sized acoustic intervals.

Data handling and analyses were performed in R (version 4.1.0), mainly using packages of the tidyverse for data wrangling and visualization (v1.3.1, Wickham *et al.*, 2019). The main analyses follow the study's primary aim, which is to determine the extent of temporal adjustments as a reaction to temporal real-time perturbation. Therefore, different prosodic units will be the focus of the analyses to shed light on timing mechanisms and their prosodic unit of control and representation. First, temporal adjustments at the perturbation site and in unperturbed segments within the target word will be examined on the sound/segment level by looking into single segment durations (section 3.3.1). After that, the perturbed sequence (CC and V) will be investigated as a whole on the syllable level with respect to the applied perturbation (section 3.3.2). Finally, perturbed and unperturbed segments within the target word will be examined on the word level by looking into word-proportional duration changes between the perturbation phases (section 3.3.3). Sections 3.3.1 and 3.3.3 will follow similar analytical strategies by examining temporal adjustments during maximum perturbation in the hold phase and then assessing continuing temporal adjustments when the perturbation is removed in the aftereffect phase. By analyzing both the hold and the aftereffect phase, we can draw conclusions about the nature of reactions, i.e., to what extent they reflect online control of ongoing speech movements (e.g., online compensation or reactive feedback control) on the one hand versus updates of motor commands for further productions (adaptation) on the other. Section 3.2 follows a different approach: The analysis on the syllable level assesses the reaction magnitude relative to the applied perturbation in the whole perturbation section (CC and V) and therefore allows to compare the Stressed with the Unstressed condition subsequently. The division of analyses into segment, syllable, and word level is expected to crucially contribute to our understanding of the temporal frame in which timing mechanisms in fluent speech are controlled and represented. For clarity, the durational changes in perturbation will always be referred to as stretched or compressed, while durational changes in speakers' production will be termed lengthened or shortened.

The uttered word's stress pattern is affected by the manipulation of duration (assumed to be the most important cue to stress). Especially in the Stressed condition, the compression of the vowel weakens the stress pattern in perception. Therefore, as a secondary aim, the interdependence of non-temporal stress markers will be examined by analyzing intensity (root-mean-square (RMS) amplitude) and spectral skewness for the vowel and the fricative, as well as the aperiodicity of the vowel. Consideration of these additional aspects is expected to add substantially to the understanding of the interdependence of stress markers (and further spectral properties) in German.

3.3.1 Temporal Adjustments on the Sound / Segment Level

The first nine baseline trials were discarded from further calculations as done previously (Oschkinat and Hoole, 2020), to avoid much variance in speaking style at the beginning of the experiment and to ensure that the baseline mean is close to the baseline value where perturbation starts in the ramp phase. Over the last 11 trials of the baseline, a mean segment duration per subject was calculated to serve as a reference for productions with regular feedback, which is depicted as the horizontal zero line in visual presentation (see, e.g., Figure 3.2).

3.3.1.1 Reaction to maximum perturbation (hold phase)

For calculations of production differences between baseline (no perturbation) and hold phase (maximum perturbation), two linear mixed models were calculated using the packages lme4 (v1.1-23; Bates *et al.*, 2015) and lmerTest (v3.1-3; Kuznetsova *et al.*, 2017). The data was separated into two datasets and two models to avoid retesting on sounds: Dataset 1 incorporated the four segments CC and V of syllable 1 and CC and V of syllable 2; dataset 2 included the five segments C1 and C2 of syllable 1, C1 and C2 of syllable 2, and syllable 3. Splitting up the data into two datasets/models emerged from the circumstance that C1 and C2 should not appear within the same model as CC since this would cause a double-testing for the incorporated segments. Treating the third syllable /nən/ as one segment mainly derived from the reduction of the syllable to a single /n/ in some productions within and across speakers.

Models were gradually incremented to best fit the variance of the data without failure of convergence. Durations were modeled as the dependent variable with phase (baseline and hold phase), segment as a concatenation of segment and syllable (e.g., CC syllable 1), and condition (Stressed vs. Unstressed) as predictors with a three-way interaction between phase, segment, and condition. With the MuMIn package, that provides tools for performing model selection and model averaging (v1.43.17; Bartoń, 2020), the random effects structure was built by calculating the explained variance of the model with the fixed factors (marginal pseudo-R-squared) and the variance explained by the model additionally including the random effects (conditional pseudo-R-squared). Intercepts and slopes for phase, segment, condition, and trial were considered as random effects of the full model. Based on the pseudo-R-squared estimation and limits of convergence, intercept and a by-subject slope for phase were finally included into the model. Backward modeling with lmerTest's step function confirmed the following model architecture (R notation), using the lmer function from the lmerTest package as estimation command:

*formula = duration ~ phase * segment * condition + (phase | Subject), data = dataset_1/2.*

The three-way interaction reflects the design of the experiment precisely. Since we applied perturbation only in the hold phase and not in the baseline, and only to particular segments varying by perturbation condition, we expect highly significant interactions between the three predictors. However, for the purposes of the study and based on our hypotheses, we will present the differences in baseline vs. hold phase per segment and per condition in detail in the following; the summary of the interactions is to be found in

Appendix C. For the second model (incorporating C1, C2, and syllable 3), backwards-modeling dropped the three-way interaction (see Appendix C, Table 3.4).

Post-hoc pairwise comparisons on significant effects between hold phase and baseline per segment and condition were performed using the emmeans package (v1.4.8; Lenth *et al.*, 2018), which computes estimated marginal means (EMMs) for the factors in the linear mixed model and comparisons or contrasts among them. The alpha-level of significance for the following model interpretations was divided by two as we retested for effects with two models (alpha = 0.025). The next section presents the changes in production by reporting the estimates provided by emmeans' pairwise comparisons sorted by perturbation condition. Along with the estimates (difference between hold phase and baseline in ms), the amount of change between the two phases in percent per segment will be reported (ratio in %). Positive estimates/ratios indicate greater durations in the hold phase relative to the baseline, while negative estimates/ratios mark shorter hold phase productions relative to the baseline. For better readability, the estimates/ratios along with the standard errors, degrees of freedom, t-ratios, and p-values are presented in Table 3.1.

For the first (perturbed) syllable in the Unstressed condition, the pairwise comparison revealed no significant temporal adjustment for CC (-1.7 ms / -0.95%) but significant compensatory lengthening for the vowel (12.0 ms / 17.29%). In the second non-perturbed syllable, both segments experienced significant lengthening relative to baseline productions (CC: 9.8 ms / 5.61%; V: 11.8 ms / 7.39%). Splitting up CC into its components, which are usually considered to be sub-segments within an affricate, showed that C1 and C2 in syllable 1 behave contrarily, whereby C1 shows a tendency for shortening (-5.0 ms / -6.37%), and C2 a tendency for lengthening (3.0 ms / 3.33%). However, in the non-perturbed syllable 2, C1 showed a non-significant tendency for lengthening (C1: 2.6 ms / 3.72%), while C2 and syllable three were significantly lengthened (C2: 6.9 ms / 6.58%; syllable 3: 14.1 ms / 4.75%).

In the Stressed condition, the first non-perturbed syllable showed no significant temporal adjustments during the hold phase compared to baseline productions for either the consonants or the vowel (CC: -3.3 ms / -1.97%; V: 1.8 ms / 2.56%). In the perturbed second syllable, no significant reaction was found to CC perturbation (1.4 ms / 0.78%), but substantial compensation with significant lengthening of the vowel in production (51.8 ms / 31.88%). Splitting up the two onset consonants into their components showed no significant temporal adjustments in the non-perturbed first syllable (C1: -2.1 ms / -2.76%; C2: -1.3 ms / -1.54%). In the second (perturbed) syllable C1 was non-significantly shortened (-5.1 ms / -7.11%), and C2 non-significantly lengthened (6.2 ms / 5.74%) causing the CC sequence as a whole to retain a stable duration throughout the experiment. The third syllable experienced non-significant lengthening (4.8 ms / 1.59%).

Table 3.1: Overview of the statistical outcome for absolute durations of the emmeans' pairwise comparisons for the two lmer models. A thick bold horizontal line separates the two models (model 1: CC /tʃ/, V /e/; model 2: C1 [t] and C2 [ʃ], and syllable 3 /nən/). Calculations present the contrast hold phase – baseline. Grey backgrounds mark segments where focal manipulation was applied. Significant p-values (alpha < 0.025) in bold. Syllable and Segment appear in two different columns for providing a better overview. However, note that in the model calculation, segment is always the concatenation of Segment (e.g., CC, and syllable, e.g. Syllable 1).

Perturbation condition	standard error	degrees of freedom (df)	Syllable	Segment	estimate (ms) (H-B)	ratio (%) ((H/B)*100-100)	t-ratio	p-value
Unstressed	2.43	106	1	CC	-1.7	-0.95	-0.68	0.495
				V	**12**	**17.29**	**4.96**	**<.0001**
			2	**CC**	**9.8**	**5.61**	**4.05**	**<.0001**
				V	**11.8**	**7.39**	**4.88**	**<.0001**
Stressed	2.42	105	1	CC	-3.3	-1.97	-1.36	0.176
				V	1.8	2.56	0.725	0.470
			2	CC	1.4	0.78	0.56	0.577
				V	**51.8**	**31.88**	**21.41**	**<.0001**
Unstressed	2.96	169	1	C1	-5.0	-6.37	-1.69	0.0924
				C2	3.0	3.33	1.028	0.306
			2	C1	2.6	3.72	0.89	0.376
				C2	**6.9**	**6.58**	**2.33**	**0.021**
			3	**/nən/**	**14.1**	**4.75**	**4.77**	**<.0001**
Stressed	2.95	166	1	C1	-2.1	-2.76	-0.73	0.468
				C2	-1.3	-1.54	-0.45	0.652
			2	C1	-5.1	-7.11	-1.71	0.088
				C2	6.2	5.74	2.11	0.036
			3	/nən/	4.8	1.59	1.62	0.108

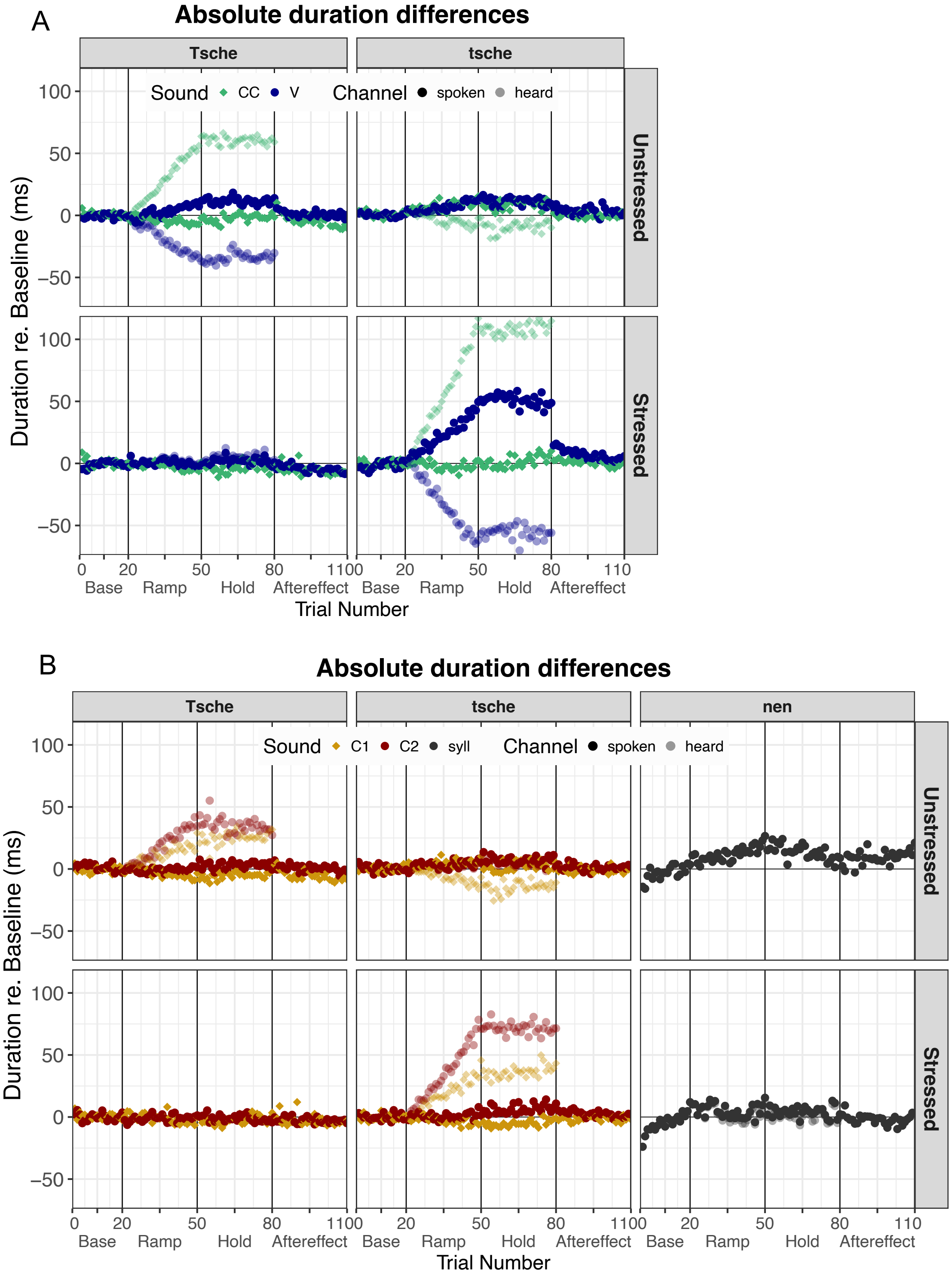


Figure 3.2: Duration differences in ms relative to the baseline mean of the onset CC (/tʃ/, green) and the vowel (/e/, blue) in syllable 1 and 2 in the upper plot **(A)**, and C1 ([t], orange) and C2 ([ʃ], red) of both syllables as well as syllable 3 (black) in the lower plot **(B)** over the course of the experiment. Solid dots mark the spoken signal, transparent dots the received perturbed auditory feedback. Unstressed condition in the upper panels per plot (perturbation of syllable 1), Stressed condition in the lower panels per plot (perturbation of syllable 2).

3.3.1.2 Adaptation – Evaluation of the aftereffect phase

General additive mixed models (GAMMs) were fitted to assess the time (or trials) over which the articulatory adjustments remain. GAMMs account for linear or non-linear relationships in the data by relying on parametric terms and smooth terms. The smooth terms define the fitted curve's shape by adding up basis functions to a more complex curve until it fits the data properly. Unlike GAMs, the mixed design incorporates random effects. Additionally to the random slope and random intercept, a random smooth parameter enables capturing by-group variation in non-linear effects (Sóskuthy, 2017; Wood, 2017). With the R packages mgcv for fitting generalized additive (mixed) models (Wood, 2011; 2017) and itsadug for evaluation, interpretation, and visualization of GAMM models (van Rij *et al.*, 2017), two models were fitted from the two datasets used for the linear mixed models: One dataset included CC and V of both syllables and conditions, the other C1 and C2 of both syllables and conditions and syllable 3. The analyses aim at visualizing the deviation of aftereffect phase productions from the baseline productions. Therefore, two curves were fitted per sound, syllable, and condition for comparison: First, a linear curve with the mean baseline duration (incorporating the last 11 baseline trials) was calculated. Secondly, the aftereffect productions were fitted. The baseline curve was stretched to 30 trials to match the aftereffect's trial numbers (trials 81 to 110). Subsequently, the difference between the baseline and the aftereffect curves of the respective segment (sound per syllable per condition) was plotted to identify regions of significant deviation (see Figure 3.3).

The GAMMS were fitted to absolute durations with the following terms: The interaction between segment and perturbation condition as a parametric term (average difference in duration depending on segment and condition); a smooth term over trial number (non-linear effect of trial number on duration) by the interaction of segment and condition; and a factor smooth which models the non-linear difference over trial number for each subject as random effect with penalty order m = 1 (to model inter-speaker variation). The primary purpose of the calculated models was to visualize statistically significant reactions over time rather than to report p-values. Statistical results would, in effect, summarize the means of the aftereffect phase and baseline. Since it is expected that reactions systematically vary within the aftereffect phase, the main interest lies in the point in time (trial number) up to which reactions diverge from the baseline mean. Visualizations of the GAMM fit illustrate the span of trials with significant effects for each segment of the word

(Figure 3.3). Confidence intervals were set to 97.5% to account for an adjusted significance level of alpha = 0.025.

The visualizations show that in the Unstressed condition (Figure 3.3.A), the vowel of syllable 1 does not differ from baseline productions. The CC segment in syllable 1 was not compensatorily shortened in the hold phase, but durations shorten from trial 91 until the end of the aftereffect phase. This is mainly caused by the significant shortening of C1 (trial 93 to 110) while C2 remains constant. The vowel in syllable 2 (the unperturbed syllable) diverges from baseline durations from trial 84 to 101. No change is seen for CC (and either C1 or C2). The third syllable, however, is significantly longer than the baseline from trial 85 to 110. In the Stressed condition (Figure 3.3.B), CC and V of syllable 1 did not change during the hold phase but start to shorten significantly when the perturbation is removed (both from trial 89 to 110). No significant change is observed for C1 and C2 and either the first or the second syllable. The vowel in syllable 2 is significantly longer from the beginning until trial 104 of the aftereffect phase. CC of syllable two and the third syllable do not diverge from baseline durations.

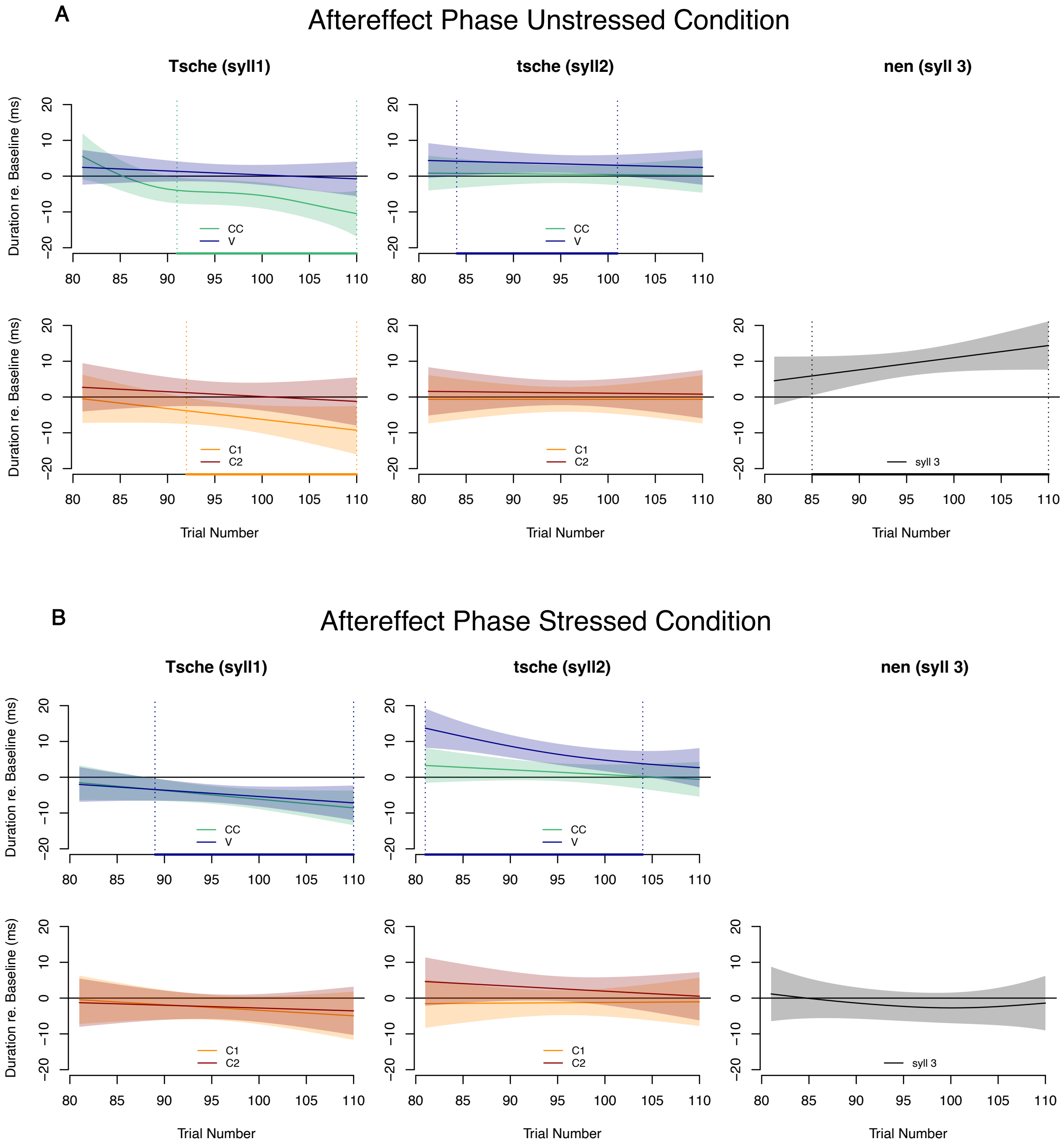


Figure 3.3: GAMM fits of the aftereffect phase for absolute durations relative to the baseline mean (ms), including random effects and confidence intervals (97.5%). The Unstressed condition visualized in the upper plot **(A)** and the Stressed condition in the lower plot **(B)** (30 subjects). The CC fits are shown in green and vowel fits are shown in blue. C1 in orange and C2 in red. The section between two dotted vertical lines and thick horizontal lines marks the significant deviation from zero for each sound.

3.3.2 Temporal Adjustments on the Syllable Level relative to the Perturbation

In the current study, it has to be taken into consideration that the vowel of the second syllable was much longer than the vowel of the first syllable due to the stress pattern. Since the perturbation section was sized to be twice the vowel duration, the perturbation section covering the stressed syllable (Stressed condition) was larger (mean: 324 ms) than the perturbation section covering the unstressed syllable (Unstressed condition, mean: 221 ms). This difference in size of the perturbation section consequently leads to a greater amount of absolute perturbation (in ms) in the Stressed condition. The following measure will take this duration difference into account by examining compensation relative to the amount of perturbation. To extract the reaction to the whole perturbed part, the following measure incorporates the segments of the whole perturbation section (CC and V) and captures the total amount of applied perturbation (stretching and compressing) to the targeted segments. This measure then gives insight into the strength of reaction relative to perturbation and allows for a comparison between the perturbation of the word-initial unstressed and the word-medial stressed syllable. Another aspect that has to be accounted for is the fact that the fit of the perturbation section changes when speakers change their productions. While the online status tracking can track the onset of the perturbation even in variable speech, the duration of the perturbation section itself, however, is not adaptive. When speakers change their productions of the perturbed segments, the location of the perturbation section may deviate from the implementation based on non-perturbed speech in the pretest. Therefore, the measurement assesses the fit of the perturbation section as compared to the baseline fit and takes into account that productions might already include compensatory/adaptive behavior.

For further analyses, the difference between baseline and hold phase productions and hold phase and (simulated) baseline perturbation will be examined to build a measure that captures the response relative to the applied perturbation. Euclidian distances of *absolute durations* between baseline and hold phase for both the produced and perceived signals will be examined.

Accordingly, two signals were considered for baseline (B) and hold phase (H), respectively: 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 perturbed signal was simulated to estimate the maximum perturbation on a signal without reaction (B2*). Example spectrograms for B1, B2*, H1, and H2 of both perturbation conditions are provided by Figure 3.1. The segments CC and V of the perturbed syllable per perturbation condition are arranged in a two-dimensional coordinate system that captures the spoken and perturbed durations of the first perturbed segment (CC, /tʃ/) on the x-axis and the spoken and perturbed durations of the second segment (V, /e/) on the y-axis (visualized in Figure 3.4. A and 3.5.B). The reference for durations is the mean baseline production (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.

A *mean perturbation* was calculated from the mean of (simulated) maximum perturbation without compensation in the baseline (Euclidian Distance |B1-B2*|, Figure 3.4.A and 3.5.B, dashed line) and perturbation on a signal that perhaps already includes a reaction in the hold phase (Euclidian distance |H1-H2|, Figure 3.4.A and 3.5.B, dashed line, see equation 1). Assuming that subjects intuitively aim to match the received auditory feedback with the representation of the intended speech sound through compensation, a closer distance between B1 (spoken and heard signal without perturbation) and H2 (heard signal/perturbed auditory feedback in the hold phase) would mean stronger compensation. If H2 equals B1, the reaction is interpreted as perfect compensation, meaning that the subjects heard the signal they intended to speak. The Euclidian distance of |B1-H2| (solid line) was then divided by the *mean perturbation* and scaled to percent values (see equation 2), forming our *compensation* values.

(1) $$mean\,perturbation = \frac{|B1-B2|+|H1-H2|}{2}$$

(2) $$compensation = 1 - \left(\frac{|B1-H2|}{mean\,pert.}\right) * 100$$

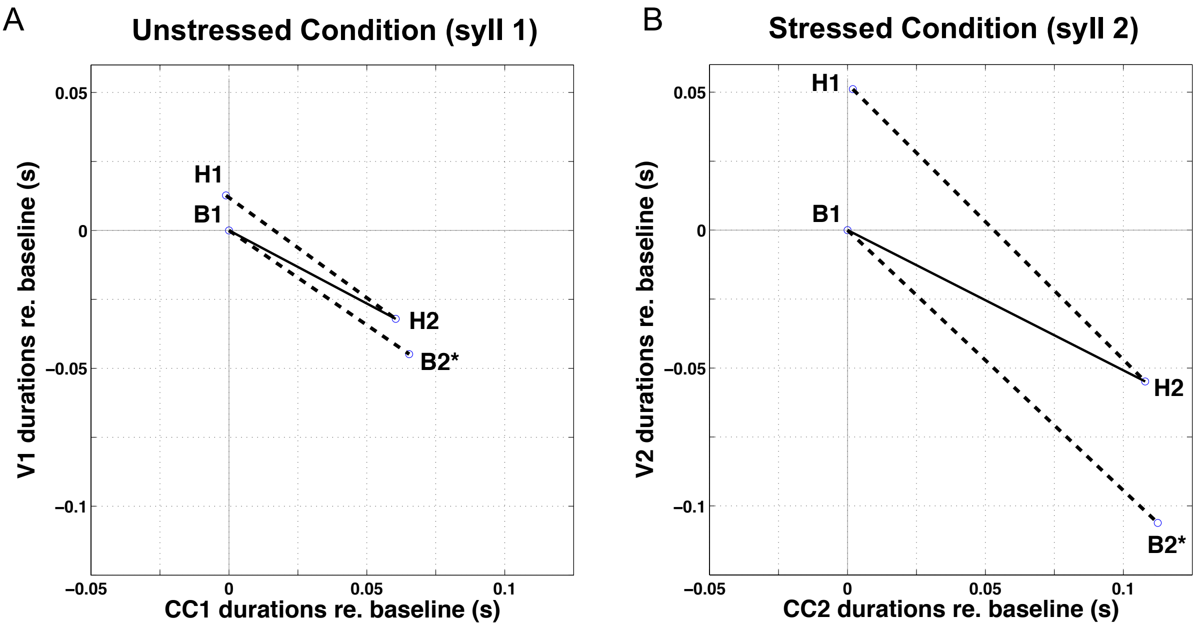


Figure 3.4: Both plots show mean durations (s) of both segments of interest (CC /tʃ/ and V /e/) over 30 subjects per perturbation condition relative to the baseline mean (0/0). The first segment of the perturbation section is on the x-axis (CC) and the second segment of the perturbation section is on the y-axis (V). 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. The left plot **(A)** shows the Unstressed condition and the right plot **(B)** the Stressed condition.

A paired t-test was fitted to compare compensation in the Unstressed condition with compensation in the Stressed condition. The outcome indicates that compensation of the whole perturbed section relative to perturbation was stronger in the Stressed condition than in the Unstressed condition (t = -2.72; df = 29, mean of the difference = -8.78, p = 0.01). Figure 3.5 visualizes the compensation magnitudes of both conditions.

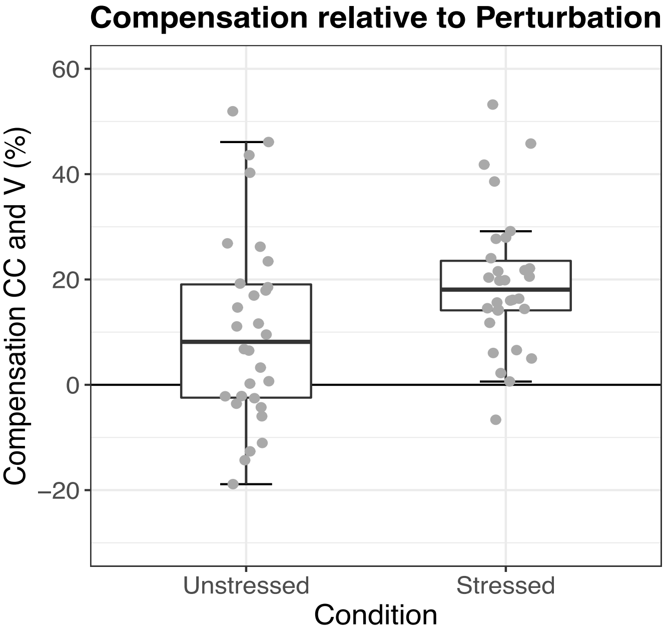


Figure 3.5: The compensation magnitude relative to perturbation for Unstressed and Stressed condition for 30 subjects. Values incorporate both perturbed segments of interest (CC and V). Boxes correspond to the first and third quartiles and bars represent the median. Whiskers extend from the hinge to the highest/smallest value no further than 1.5 IQR. Data beyond the whiskers are outliers. Dots mark individual subjects.

3.3.3 *Temporal Adjustments on the Word Level*

To estimate the impact of durational adjustment in a higher prosodic unit, all absolute durations were normalized by word duration of the respective trial (% of word duration). The following analyses reveal how the proportional segment durations within the word change over the course of the experiment.

3.3.3.1 *Reaction to maximum perturbation (hold phase)*

For calculations of production differences between baseline (no perturbation) and hold phase (maximum perturbation), two linear mixed models with the same structure as in section 3.3.1.1 were calculated but with *normalized durations* as the dependent variable. Accordingly, as above, the alpha-level of significance for the following model interpretations was divided by two as we retested for effects with two models (alpha = 0.025). The following section reports the estimates provided by *emmeans*' pairwise comparisons; Table 3.2 summarizes more details of the outcome. Along with the *estimate* that reports the difference of proportion of a segment in the word between baseline and hold phase (H-B in %), we report the *ratio* as the change in word-normalized segment duration of the respective segment between baseline and hold phase ((H/B * 100) – 100 in %), the latter reported in Oschkinat and Hoole (2020). For example, an *estimate* of 25% means that the segment takes up 25% more of the word in the hold phase than in the baseline. A *ratio* of 25% indicates that the word-normalized segment is 25% longer in the hold phase than in the baseline. Figure 3.6 depicts the duration of a segment within the word relative to the calculated baseline mean per segment.

The models' outcomes for the Unstressed condition showed significant shortening of CC in the first syllable (estimate: -1.17%; ratio: -6.03%) and significant lengthening of the vowel (estimate: 0.88%; ratio: 11.03%). In the second non-perturbed syllable, the CC segment did not change significantly in production (estimate: 0%; ratio: 0%), while the vowel was significantly lengthened (estimate: 0.47%; ratio: 2.51%). Splitting up CC into its components showed that C1 and C2 in syllable 1 were both shortened, C1 significantly (estimate: -1.05; ratio: -11.38%), C2 non-significantly (estimate: -0.21%; ratio: -1.92%). In the non-perturbed syllable 2, both consonants did not change significantly (C1 estimate: -0.15%; ratio: -1.84%; C2 estimate: 0.07%; ratio: 0.57%). The third syllable did not show a significant change in duration (estimate: -0.14%; ratio: -0.41%).

In the Stressed condition, the CC segment of the first non-perturbed syllable was significantly shorter than baseline productions (estimate: -1.45%; ratio: -7.62%), but the vowel did not change significantly (estimate: -0.18%; ratio: -2.25%). In the perturbed second syllable, both CC and V show significant compensatory temporal adjustments (CC estimate: -1.19%; ratio: -5.8%; V estimate: 4.38%; ratio: 23.84%). C1 and C2 were both significantly shortened in syllable 1 (C1 estimate: -0.68%; ratio: -7.89%; C2 estimate: -0.74; ratio: -7.12%), while in syllable 2 C1 was significantly shortened (estimate: -0.99; ratio: -12.12%) but C2 rather remained constant (estimate: -0.18%; ratio: -1.38). The third syllable was significantly shorter than in the baseline (estimate: -1.55; ratio: -4.57%).

Figure 3.6 visualizes the durational adjustments throughout the experiment for each segment of interest relative to the baseline mean (horizontal zero lines).

Table 3.2: Overview of the statistical outcome for normalized durations of the emmeans' pairwise comparisons for the two lmer models. A thick bold horizontal line separates the two models (model 1: CC /tʃ/, V /e/, model 2: C1 [t], C2 [ʃ], and syllable 3 /nən/). Grey backgrounds mark segments that were perturbed. Significant p-values (alpha < 0.025) in bold. Syllable and Segment appear in two different columns for providing a better overview. However, note that in the model calculation, segment is always the concatenation of Segment (e.g., CC, and syllable, e.g. Syllable 1).

Perturbation condition	standard error	degrees of freedom (df)	Syllable	Segment	estimate (%) of word (H – B)	ratio (%) re. baseline ((H/B)*100 -100)	t-ratio	p-value
Unstressed	0.199	338	1	**CC**	**-1.17**	**-6.03**	**-5.91**	**<.0001**
				V	**0.88**	**11.03**	**4.41**	**<.0001**
			2	CC	0.00	0.00	-0.01	0.996
				V	**0.47**	**2.51**	**2.34**	**0.020**
Stressed	0.198	331	1	**CC**	**-1.45**	**-7.62**	**-7.32**	**<.0001**
				V	-0.18	-2.25	-0.92	0.357
			2	**CC**	**-1.19**	**-5.80**	**-6.02**	**<.0001**
				V	**4.38**	**23.84**	**22.17**	**<.0001**
Unstressed	0.203	1769	1	**C1**	**-1.05**	**-11.38**	**-5.16**	**<.0001**
				C2	-0.21	-1.92	-1.01	0.310
			2	C1	-0.15	-1.84	-0.73	0.468
				C2	0.07	0.57	0.32	0.745
			3	**/nən/**	-0.14	-0.41	-0.67	0.505
Stressed	0.202	1734	1	**C1**	**-0.68**	**-7.89**	**-3.36**	**0.001**
				C2	**-0.74**	**-7.12**	**-3.67**	**<.0001**
			2	**C1**	**-0.99**	**-12.12**	**-4.89**	**<.0001**
				C2	-0.18	-1.38	-0.87	0.382
			3	**/nən/**	**-1.55**	**-4.57**	**-7.70**	**<.0001**

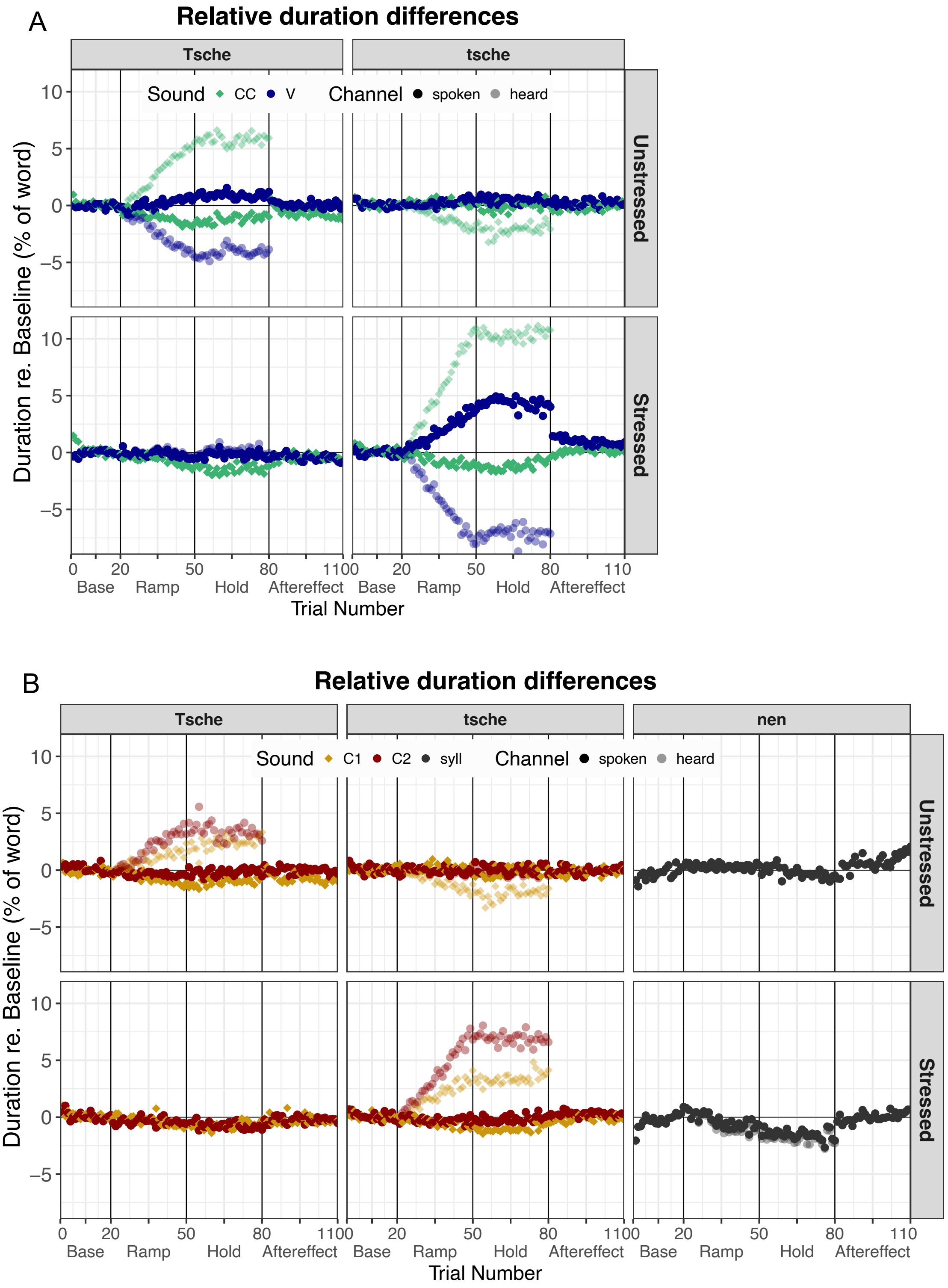


Figure 3.6: Duration differences relative to the baseline mean (estimate in %) of the onset CC (/tʃ/, green) and the vowel (/e/, blue) in syllable 1 and 2 in the upper plot **(A)**, and C1 ([t], orange) and C2 ([ʃ], red) of both syllables as well as syllable 3 (black) in the lower plot **(B)** over the course of the experiment. Solid dots mark the spoken signal, transparent dots the received perturbed auditory feedback. Unstressed condition in the upper panels (perturbation of syllable 1), Stressed condition in the lower panels (perturbation of syllable 2).

3.3.3.2 Adaptation – Evaluation of the aftereffect phase

Similar to the analyses of absolute durations in section 3.3.1.2, general additive mixed models (GAMMs) were fitted with the same model structure as described in section 3.3.1.2, but to normalized durations to assess for how many trials of the aftereffect phase the articulatory adjustments remained.

In the following, the outcome given by the visualization of the Gamms will be reported. Figure 3.7.A indicates that in the Unstressed condition, the lengthening of the vowel in the first syllable did not continue in the aftereffect phase, while the CC segment was significantly shortened from trial 87 to trial 110. The shortening was mainly caused by C1 (significant deviation from trial 87 to 110), while C2 maintained baseline durations. In syllable two, no significant effects were found. Syllable three was longer than the baseline from trial 88 to trial 110. In the Stressed condition (Figure 3.7.B), CC and V of syllable 1 were significantly shorter than baseline productions (CC: trial 84 to 110, V: trial 88 to 110), C1 and C2 did not diverge significantly. In syllable 2, the vowel was significantly longer from trial 81 to 110 (comprising the whole aftereffect phase), while CC remained constant. No significant effect was found for C1 or C2 in syllable 2 or syllable 3.

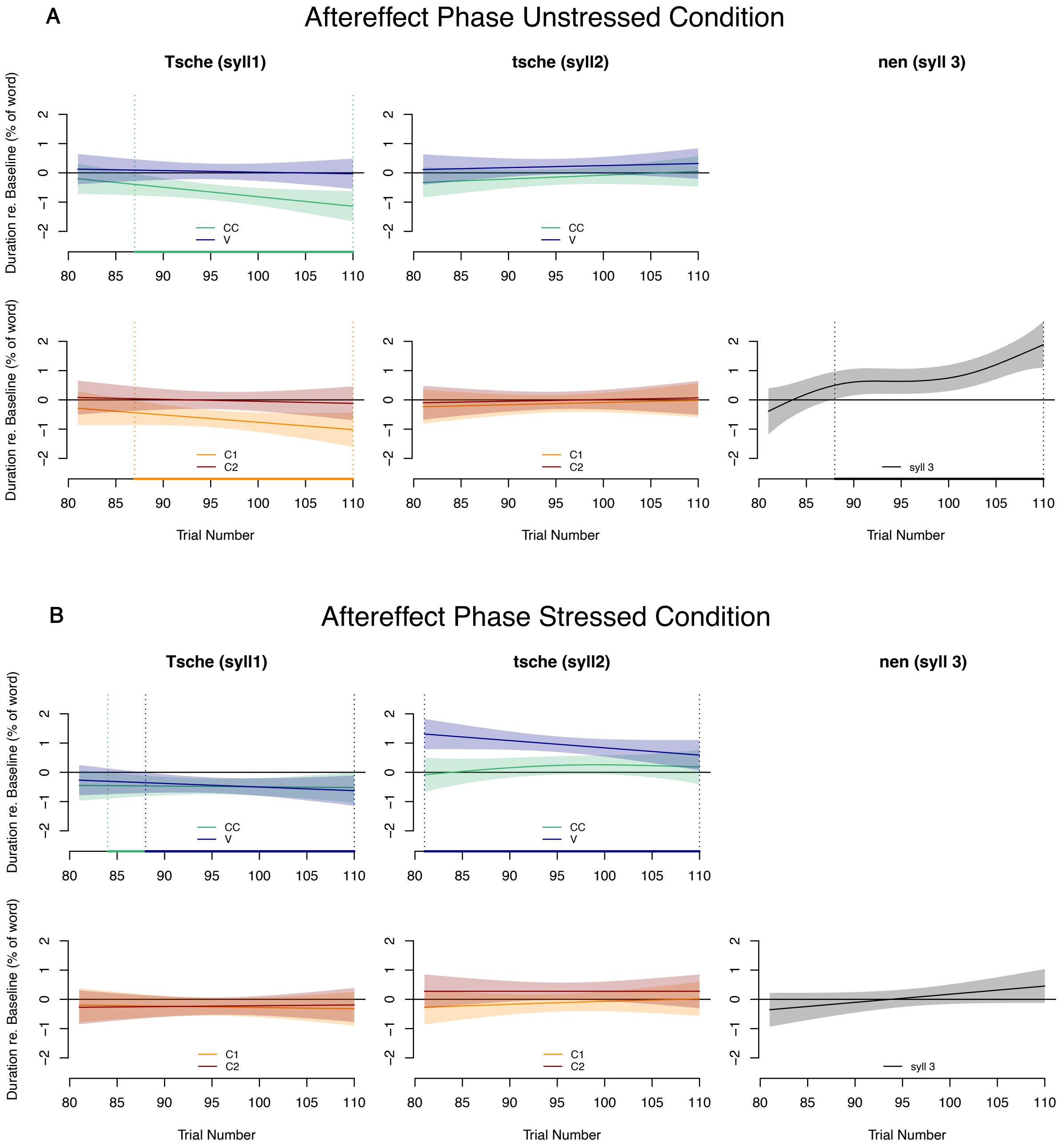


Figure 3.7: GAMM fits of the aftereffect phase for word normalized durations relative to the baseline mean (%), including random effects and confidence intervals (97.5%). The Unstressed condition in the upper plot **(A)** and the Stressed condition in the lower plot **(B)** (30 subjects). The CC fits are shown in green and vowel fits are shown in blue. C1 in orange and C2 in red. The section between two dotted vertical lines and thick horizontal lines marks the significant deviation from zero for each sound.

3.3.4 Non-temporal Markers of Stress

The temporal perturbation in this study compressed the vowel in both perturbation conditions. Consequently, in the Stressed condition, the stress pattern was attenuated. It is therefore assumable that not exclusively duration but also other markers of stress may have changed in production. The following sections examine aperiodicity of the vowels in both perturbed syllables, intensity (root-mean-square amplitude), and spectral skewness for the vowel and the fricative. The fricative will also be examined to reveal possible production differences in change of intensity (RMS) and skewness of the spectrum. Please recall that only the first syllable was perturbed in the Unstressed condition, while in the Stressed condition, the second syllable was perturbed.

As a reminder, we expect more intensity (RMS) in the perturbed vowel or fricative, less skewness in the perturbed vowel or fricative, and less aperiodicity in the perturbed vowel. Since we observed greater absolute durational adjustments in the vowel of the Stressed condition than in the Unstressed condition, we expect changes in intensity, skewness, or aperiodicity to be more pronounced in the Unstressed condition. All calculations and visualizations incorporate the last ten trials of the baseline exclusively. In visualization, the aftereffect phase is added for an overview; calculations include baseline and hold phase exclusively. The examination of the mentioned parameters is rather a secondary aim of the study and should be seen as exploratory in nature. Therefore, we retain unadjusted p-values in the following and ask the reader to keep that in mind when interpreting the following outcomes.

3.3.4.1 Aperiodicity

Aperiodicity was estimated with the Matlab function yin (Cheveigné and Kawahara, 2002). The mean aperiodicity values for each vowel segment were entered into the analyses below.

Aperiodicity values were provided by yin on a scale between 0 and 1. The data were not normally distributed and consequently log-transformed for calculations and plots. More strongly negative values (after transformation) indicate less aperiodicity, while smaller negative values reflect greater aperiodicity. Values were grouped by sex, condition, and phase and all values outside the 95% confidence intervals were removed. The left panel of

Figure 3.8 shows log-transformed aperiodicity values per condition and sex, the right panel presents the log-transformed aperiodicity values normalized by each subject's baseline mean per condition. A linear mixed model was fitted with log-transformed aperiodicity values as the dependent variable with phase, condition, and sex as predictors and an interaction between phase and condition. The interaction between phase and condition was added as a within-subject random effect (intercept and slope). Emmeans' comparison between the Unstressed condition (syllable 1) and the Stressed condition (syllable 2) averaged over phase and sex indicated that the vowel in the unstressed syllable was produced with greater aperiodicity than the stressed vowel (estimate syll1-syll2: 0.97; SE = 0.064; df = 29; t-ratio =15.08; p <.001). Further, the comparison between male and female subjects revealed less aperiodicity for male subjects averaged over phase and condition (female-male estimate = -1.3; SE = 0.108; df = 28; t-ratio = -12.026; p <.001). The pairwise comparison between the phases revealed significantly less aperiodicity in the hold phase compared to the baseline in the Unstressed condition (H-B estimate = -0.096; SE = 0.039; df = 29; t-ratio = -2.403; p = 0.0229) and in the Stressed condition (H-B estimate = -0.177; SE = 0.039; df = 29; t-ratio = -4.531; p < .001).

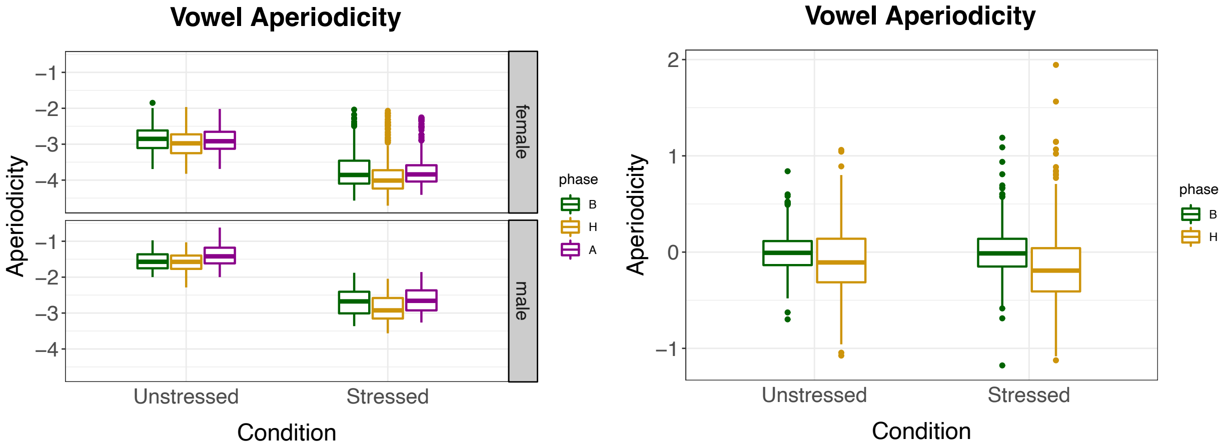


Figure 3.8: Aperiodicity (log-transformed) of the vowels in both perturbation conditions split by sex (left panel). The right panel shows aperiodicity values (log-transformed) relative to the baseline mean for baseline and hold phase. Boxes correspond to the first and third quartiles and bars represent 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. Boxplot statistics apply to all following boxplots.

3.3.4.2 *Root-mean-square of the amplitude of the signal (RMS)*

The RMS values were extracted as given by the Audapter software as an average across the entire segment (the fricative and the vowel). Data were not normally distributed and subsequently log-transformed. For the reduction of measuring errors, data were grouped by segment, condition, and phase and data beyond the 95% confidence intervals were removed. Greater negative values indicate less intensity, smaller negative values greater intensity. Figure 3.9 shows log-transformed RMS values grouped by perturbation condition and segment for each phase of interest.

A linear mixed model was calculated with log-transformed RMS values as the dependent variable with phase, perturbation condition, and segment as predictors with an interaction between phase and condition and segment. A by-subject interaction between phase and perturbation condition was added as a random effect.

Post-hoc testing with *emmeans'* pairwise comparison indicated significantly more intensity in the hold phase than in the baseline in the Unstressed condition for the vowel (estimate = 0.149; SE = 0.04; df = 36.4; *t-ratio* =3.747; *p* < .001) but not for the fricative (estimate = 0.0569; SE = 0.0398; df = 36.3; *t-ratio* =1.43; *p* = 0.161). In the Stressed condition, both segments were produced with greater intensity (vowel estimate = 0.103; SE = 0.034; df = 37.8; *t-ratio* =3.747; *p* < .001, fricative estimate = 0.126; SE = 0.0367; df = 37.8; *t-ratio* =3.452; *p* = 0.0014).

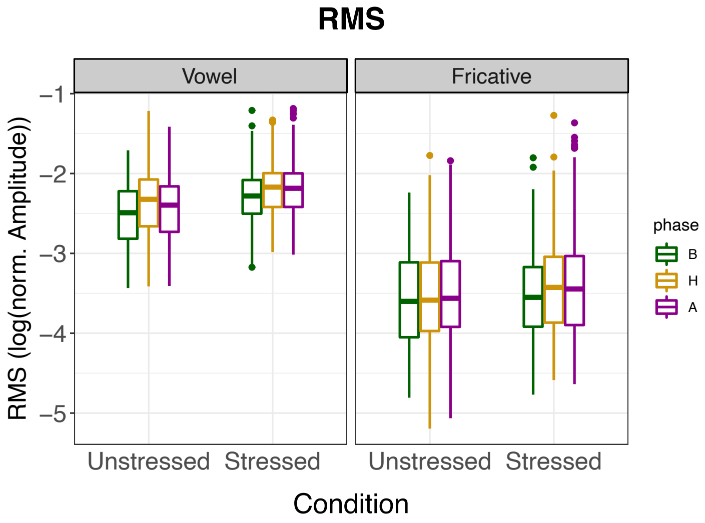


Figure 3.9: log-transformed and scaled RMS values (y-axis) in both perturbation conditions split by segment (vowel/fricative).

3.3.4.3 Skewness of the spectrum

The last examined parameter was spectral skewness. The skewness captures whether the shape of the spectrum below the center of gravity is different from the shape above the center of gravity and whether this relation changes in the face of the perturbation. For the estimation of the skewness, the standardized 3rd moment of the spectrum was extracted in the inner 50% of the sound, with a minimum duration of 240 samples (15ms) for the vowel or 320 samples (20ms) for the fricative /ʃ/. For the calculations within the fricative, frequencies between 800 and 8000 Hz were extracted with a sample rate of 16000 Hz. For the vowel, frequencies between 70 and 4000 were extracted with the same sample rate. Data outside the 95% confidence intervals were removed. A higher skewness value indicates more energy in lower frequencies than in higher frequencies.

The vowel spectra had a positive skew (mean: 4.46, range: -0.6 to 18.6), and the fricative spectra were mostly positive but for some speakers negatively skewed (mean: 0.57, range: -1.38 to 5.17). Figure 3.10 gives an example of spectral shape for the vowel and the fricative of one (male) speaker with a skewness of 10.6 for the vowel and 0.8 for the fricative.

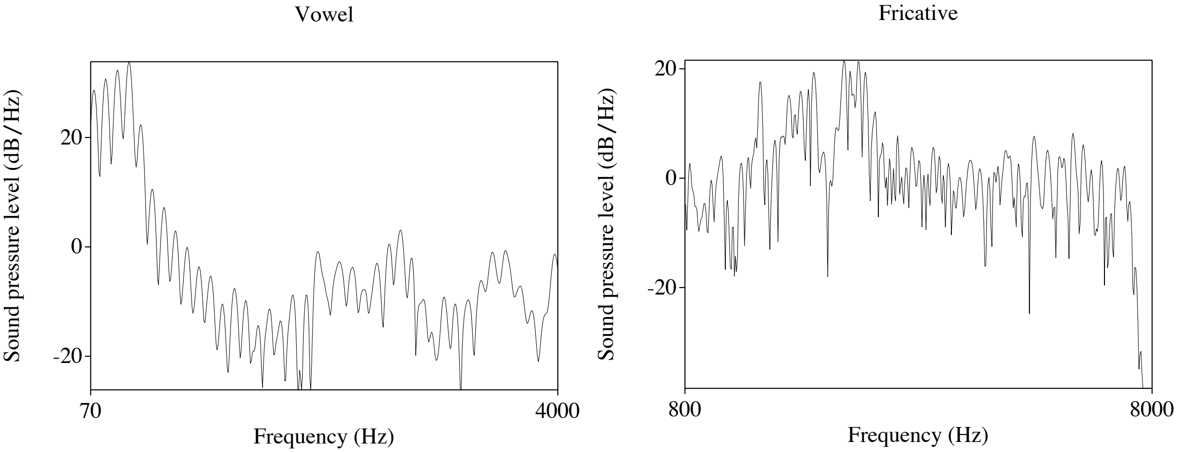


Figure 3.10: Spectral slices of the vowel /e/ (left panel) and the fricative /ʃ/ (right panel) in the second syllable of the word "Tschetschenen" spoken by a male speaker. Measures were taken in the inner 50% of the sounds in a baseline trial. Both spectra are positively skewed.

A linear mixed model was calculated with similar structure as before: skewness was the dependent variable, with phase and segment and condition as predictors with an interaction between phase and segment and condition. The interaction between phase and condition was added as within-subject random effect. Post-hoc testing revealed a significant difference between baseline and hold phase for the vowel in both conditions with less skewed spectra in the hold phase (Unstressed condition: estimate = -0.357; SE = 0.106; df = 68.7; *t-ratio* = -3.382; *p* = 0.0012; Stressed condition: estimate = -0.503; SE = 0.1; df = 77.4; *t-ratio* = -5.036; *p* <.001). No difference was observed in the fricative spectral tilt for either condition (Unstressed condition: estimate = -0.049; SE = 0.105, df = 66.3; *t-ratio* = -0.466; *p* = 0.64, Stressed condition: estimate = 0.042; SE = 0.099; df = 73.4; *t-ratio* = -0.422; *p* = 0.67). Figure 3.11 visualizes spectral skewness of the vowel and the fricative in both perturbation conditions.

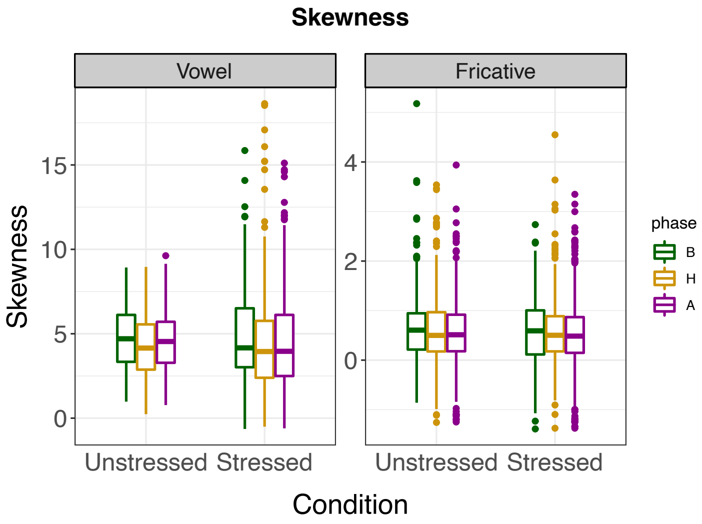


Figure 3.11: Skewness (y-axis) in both perturbation conditions split by segment (vowel/fricative).

3.3.4.4 *Interdependence of parameters and summary*

To test for dependencies of parameters, intensity, skewness, and aperiodicity were correlated with each other. For this calculation, the difference between mean values for baseline and hold phase (H-B) were estimated per speaker and condition for the vowel in the perturbed syllable. Linear models per condition per two of the above parameters were calculated. For the vowel in the Unstressed condition (syllable 1) the model revealed a significant change of aperiodicity along with intensity (RMS), whereby aperiodicity decreases with higher intensity in the hold phase (F-statistic: 10.15, DF: 28, p: 0.0035). The remaining models showed no significant effect.

Before turning to the discussion, we briefly summarize the previous section by noting that along with greater duration of the vowels in both conditions their intensity increased, their spectrum became less aperiodic and less skewed. The fricative /ʃ/ only experienced more intensity in the Stressed condition along with greater duration.
Accordingly, there are no between-condition differences that indicate a systematic contribution of stress pattern (stressed or unstressed syllable) to the responses.

3.4 Discussion

The current study revealed speakers' sensitivity to temporal manipulations in both a stressed and an unstressed syllable. This effect has been shown before, albeit only very recently, for perturbations of stressed and unstressed syllables in the spectral domain (Bakst and Niziolek, 2021). Thus, the present study contributes to a better understanding of whether processing patterns found in response to real-time spectral alterations extend to the less explored but clearly equally crucial area of real-time temporal alterations. In the current study we observed local compensatory behavior in the perturbed sequences and elicited different systematic response strategies for the global control of higher prosodic timing dependent on stress pattern (and syllable position).

We first consider absolute durations as presented in section 3.3.1. It turns out the patterns found there lead very naturally into a discussion of relative durations at the word level (section 3.3.3). Following that we return to a consideration of syllable level effects and the comparison of both perturbation conditions. Subsequently, we interpret the adaptive behavior as well as the results for non-temporal parameters.

3.4.1 Duration and Timing during Perturbation

3.4.1.1 Compensation on segment level

On the sound/segment level, speakers reacted as expected in both perturbed syllables: They significantly compensated for the auditorily compressed vowel /e/ by lengthening it in production but did not compensate significantly for the stretched CC onset segment taken as a whole. Adjustments to the single onset consonants in the perturbed syllable were also non-significant (except for C2 in the Unstressed condition), but showed a pattern in directionality for C1 [t] to shorten and C2 [ʃ] to lengthen in production. This pattern might be a result of sound class specific production and intelligibility: While the approximation of the closure of a plosive (as is C1) is sufficient to make it perceivable as a plosive, producing a fricative (as is C2) requires greater precision in building the fricative-specific constriction and a minimum duration. However, the different response directionality could support the idea that both single consonants are timed individually rather than as one single unit (affricate).

The above findings are in line with our previous study (Oschkinat and Hoole, 2020), where perturbations of the onset /pf/ and nucleus /a/ led to a non-significant shortening of the initial plosive [p] (which was a compensatory response) and non-significant lengthening of the second consonant [f] (following the perturbation). These tendencies resulted in no change in production of the whole CC /pf/ onset segment. For the compressed /a/ in manipulation, subjects compensated significantly. While in the current study the responses at the perturbation site itself are pretty similar in both perturbation conditions (Unstressed/Stressed), the temporal re-organization of unperturbed parts differs remarkably.

In the Stressed condition, the vowel of the perturbed stressed syllable was lengthened in production, and indeed very substantially (mean 51.8 ms), while the other segments within the word kept a constant duration. Since this is the stressed syllable, we hypothesize that the vowel has a critical limit on how short it can be but no strict limit on how long it can be (in contrast to the vowel in the first syllable). In the Unstressed condition, the word-initial manipulation caused global lengthening in production for all following segments in syllable two and syllable three. This reaction is reminiscent of Cai *et al.* (2011), who found lengthening of segments in the immediately following syllable after perturbation as a response to a delayed vowel target. Like the reactions in Cai *et al.* (2011), our data call to mind effects of delayed auditory feedback, which include prolongations or slowing down of following segments (Yates, 1963). The stretching of the onset consonants in perturbation caused a delay of the vowel onset which might have triggered prolongations in the following syllables. The following perturbatory compression of the vowel, which brought the signal back to real-time again, seemed to have only minor repercussions. Some of our subjects developed stutter-like symptoms during the perturbation by repeating the third syllable (see section 3.2.4), which is another indication for a reaction caused by delayed auditory feedback. In some cases, variability in production caused variability in perturbation timing, which in in turn led in some cases to compression not only for the vowel of syllable one but also of the CC segment in the second syllable. This compression of CC in syllable two might have enhanced lengthening responses. However, we assume global lengthening would be the same even without the spill-over manipulation to the second syllable.

Why does the temporal perturbation of a word-initial, unstressed syllable cause a global reaction of timing, while the temporal perturbation of a word-medial, stressed syllable just elicits local adjustments of vowel duration? We conclude that the perturbation triggers different timing strategies to maintain a higher prosodic target that are, as we assume, shaped by both the position and the stress pattern of the perturbed syllable.

If the first syllable or the onset is manipulated, so that it is perceived longer/slowed down, the timing in the higher prosodic unit (syllable/word) can be adjusted dynamically with adjustments of the following segment durations. With no shortening in production of the CC segment but lengthening of the vowel in the unstressed first syllable, the whole first syllable is longer than before, and the following adjustments aim at matching the appropriate proportional duration of each syllable within the word. Accordingly, the perturbation of the word-initial syllable might have triggered the perception of a general speech rate shift. In the perturbation of the second, stressed syllable (Stressed condition), only the vowel in the stressed syllable was perceived as being too short and consequently the marking of the vowel as stressed seemed to be of highest priority. In our data, the same technical perturbation leads to different timing strategies (global maintenance of speech rate or local adjustments to mark the stress pattern), indicating that the perception of the same shift might differ depended on where it is applied. As for the Stressed condition it also has to be kept in mind that the stressed syllable also carries the phrasal accent which cumulates in a high prominence on the stressed/accented syllable. The accentuation might lead to intensified hyperarticulation (De Jong, 1995; Cho, 2009; Mücke and Grice, 2014) when the stress/accent pattern is attenuated in the auditory feedback during perturbation. From a phonological perspective, Saltzman *et al.* (2008) introduced the μ-gesture as a temporal modulation gesture to create appropriate durational differences between stressed and unstressed syllables. The μ-gesture slows the stressed syllable down, while the duration of unstressed syllables in a foot (syllable 3 in "Tschetschenen") does not change. The response patterns in the Stressed condition in this study seem to support the idea of the μ-gesture as a function for localized slowing down of the stressed syllable.

3.4.1.2 Compensation on Word-level

The global lengthening in production of segments in the Unstressed condition during perturbation paints a clear picture of the word level's timing strategy when viewed in word-normalized durations: All segments from the vowel in the first syllable onwards were lengthened (in absolute durations), which leads to a proportionally shorter CC segment in syllable one. The perturbed unstressed vowel in syllable one is proportionally longer when viewed on word-level, while all following segments take up as much in the word as without perturbation (see Figure 3.6).

In the Stressed condition on the other hand, the unperturbed first syllable did not experience significant temporal adjustments in production, and neither did syllable three. Both unperturbed syllables maintained a stable production duration throughout the experiment. However, due to the strong compensatory lengthening of the vowel in the medial perturbed stressed syllable (51.8 ms), the other segments within the word take up less space in the word than they did in the baseline (CC in syllable one and two, and syllable three). This effect leads to the suggestion of a compensatory shortening for CC in the perturbed stressed syllable in word-normalized durations.

In summary, we conclude that on the sound/segment level, local compensation is only found for the vowel in both conditions. On the word level, however, speakers compensate bidirectionally (with compensatory lengthening and compensatory shortening) for both perturbed segments (V and CC) (achieving this aim with adjustments of following segments in the Unstressed condition). This interpretation leads us to a more differentiated use of terminology: While on the segment level, speakers compensate for the sound-specific *duration,* adjustments on the word level indicate compensation in *timing* and *coordination* of single sound durations within a higher prosodic unit.

This terminology aims at capturing different levels of processing and organization with respect to the temporal properties of speech; it reflects ideas that have been entertained about the spatiotemporal properties of phonological gestures. For example, these have been suggested to contain a spatial dimension (spectral or constriction target) and two timing dimensions: internal timing (durational properties on a segmental level) and inter-gestural timing (coordination of gestures within higher prosodic structures, Byrd and Choi, 2010).

3.4.1.3 Comparison of the Stressed and Unstressed Condition (Syllable Level)

In comparing both perturbation conditions, we expected greater compensation to the stressed vowel since the perturbation auditorily weakened the desired stress pattern. A counteraction to the perturbation would maintain the desired stress pattern of the word. The production difference for the vowel /e/ was much more substantial in the perturbed stressed syllable (51.8 ms) than in the perturbed unstressed syllable (12 ms). However, the stressed vowel in the second syllable was also much longer than the unstressed vowel in the first syllable, and therefore the perturbation was greater in the stressed syllable. The calculations on the syllable level in section 3.3.2 incorporated the whole perturbation section (CC and V) and the amount of perturbation. The results indicated greater compensatory responses to the stressed, second syllable than compensation to the first, unstressed syllable. This outcome supports our hypothesis that speakers aim at realizing the intended lexical stress pattern by adjusting the duration of the stressed syllable to a greater extent than compensating for the unstressed syllable. Taking the whole preceding discussion into account, however, this result has to be interpreted cautiously since we showed that compensation to the perturbed first syllable in the Unstressed condition was not exclusively realized in the first syllable, but also spread over the whole word.

Admittedly, adjustments in unperturbed syllables were not captured in the analyses at the syllable level in section 3.3.2.

Moreover, one aspect that we cannot rule out concerns the different syllable positions in both perturbed sequences. While it is likely that the stress pattern causes the more robust response in the perturbed syllable, it can additionally or as an alternative be caused by the fact that the stressed syllable appears word-medially while the unstressed syllable is word-initial, the former having more temporal context information available for word timing than the latter. Syllable position was found to affect reactions to (supra)segmental spectral alterations in previous studies, with a complex interaction with stress pattern (Natke and Kalveram, 2001; Bakst and Niziolek, 2021).

3.4.2 *Compensation, Adaptation, and Reactive Feedback Control*

While we have noticed different global reaction patterns between the two perturbation conditions, the response's nature is not entirely characterized by exclusively observing the hold phase productions. The analyses of the aftereffect phase allow differentiation as to whether the feedforward representation for production was updated or whether online control drove changes in the ongoing trial itself.

In the Unstressed condition, CC of the perturbed first syllable is shortened in production in the aftereffect phase (in absolute and word-normalized durations). This reaction might follow the aim of keeping the vowel relatively long compared to the CC segment when the vowel itself is not produced longer anymore. Similarly, CC and the vowel of the unperturbed first syllable in the Stressed condition are both produced shorter in the aftereffect phase (with a faster speech rate), to make the second syllable sound more stressed (in absolute and word-normalized durations). In this view, the systematic aftereffects aim at keeping the established relation between CC and V in the Unstressed condition and between syllable one and syllable two in the Stressed condition, but by changing segments other than the initially perturbed parts. This response pattern additionally indicates that the onset in general can in fact be adjusted in production, but perhaps not as a reaction to locally applied perturbation, but rather caused by a mismatch in timing with other segments in the syllable/word.

In the planning and control of timing, the first syllable might set the temporal grid for the following syllables within a word, forming a counterpart to our proposal that the onset sets a grid for following sound durations within the syllable (Oschkinat and Hoole, 2020). This interpretation is in line with the perception study by Reinisch *et al.* (2011), who tested the perception of stress in different syllable positions dependent on speech rate. When the initial syllable was slowed down, the second syllable sounded shorter and therefore unstressed. Reinisch *et al.* (2011) further concluded that judgments about the stress pattern are made on initial syllable duration, regardless of the stressed syllable's position within the word. This conclusion is closely related to concepts in spoken-word recognition, where the listener uses information as soon as it is available for decoding and word-recognition (e.g., Reinisch *et al.*, 2010; 2011). The systematic aftereffects in the first syllable in both conditions suggest that in perception and production speakers aim to provide as much information as possible as early as possible, which complicates the attribution of specific cues to purely production or perceptual mechanisms. Whether or not the responses in the aftereffect phase can be seen as adaptive depends on the reaction in the hold phase: Responses in the aftereffect phase that remain similar to the responses in the hold phase indicate adaptive behavior, further aftereffect responses that deviate from hold phase and baseline productions indicate a reactive feedback response to the withdrawal of feedback shift, with the aim to keep the relation between segments within the syllable or syllables within the word constant.

Adaptive responses are seen in the Unstressed condition in the vowel of the second syllable and syllable three. While the vowel in the second syllable has probably updated its durational target towards longer durations to mark the stress pattern, we admittedly have no explanation nor assumption for why syllable three also adapts towards longer durations.

In the Stressed condition, the vowel in syllable two experiences strong adaptive behavior. However, there is a noticeably large drop from the end of the hold phase to the beginning of the aftereffect phase (see Figures 3.2 and 3.6). While there is substantial compensation during the whole hold phase, with the first trial of the aftereffect phase, the vowel shortens abruptly. This behavior indicates a strong component of within-trial reactive responses (online compensation) to the ongoing perturbation in the hold phase. The actual amount of update in the motor commands is indicated by the starting point of durations in the aftereffect phase, while the size of the drop from the hold to the aftereffect phase indicates the additional online compensation component. However, online compensation is only possible with lengthening of segments since it is impossible to shorten segments in real-time as a reaction to a longer percept. Lengthening the vowel in the online control might also be driven by the circumstance that the first segment (CC) is stretched in perturbation, and lengthening of the second segment (V) in production also compensates for the first segment when viewed from the perspective of larger timing units.

Comparing both conditions indicates that the global lengthening of segments in the Unstressed condition is mainly indicative of online control mechanisms in an ongoing speech sequence. In contrast, the systematic adjustments to the first syllable and the vowel of the second stressed syllable in the aftereffect phase in both conditions indicate an update of the motor commands for the relation between stressed and unstressed syllable within the word.

The current study's paradigm allowed the examination of adaptation effects from the hold phase to the aftereffect phase and transfer of adaptation effects from one perturbed syllable to a similar non-perturbed syllable. Our data suggest no adaptation effects in within-trial moment-to-moment control from the perturbed word-initial to the non-perturbed word-medial syllable in the Unstressed condition: Even though the segments in the second syllable of the Unstressed condition are lengthened in production, this does not necessarily reflect transmission of compensatory behavior from the first syllable to the second, but indicates a general slowing down. In between-trial transmission from the perturbed word-medial to the unperturbed word-initial syllable in the Stressed condition, we do not see effects in absolute durations (on the segment level). On the word level (in word-normalized durations), the CC segment in the first (unperturbed) syllable is shortened to the same degree as CC in the second (perturbed) syllable. This, however, is not directly attributable to a transmission of compensatory response from the second to the first syllable, since all segments appear shorter due to the lengthened vowel in the second syllable (as discussed above). Further, the vowel in the first syllable does not change remarkably.

However, in spectral perturbation studies, effects of transmission from a perturbed vowel to the same vowel in another word have been observed. Houde and Jordan (2002) found learning effects due to compensation of the vowel in a CVC word partially transferred to another CVC word with the same vowel, suggesting that the vowels in both words share the same representation. Caudrelier *et al.* (2018) further tested transfer of vowel adaptation from the perturbed monosyllabic /be/ to the unperturbed pseudowords /bepe/, /pebe/, and the real-word /bebe/. Their participants transferred learned production updates but with greater transfer to the same syllable /be/ than the similar one /pe/ and greater transfer to the first than the second syllable. The lack of adaptation transfer in our data raises the question of whether segment duration and syllable timing share the same representation when they appear in different syllables and the syllables in a different position within the word. Our findings from this study suggest that the temporal control depends on stress pattern, syllable position within the word, and, as previously shown, segment position within the syllable (Oschkinat and Hoole, 2020).

3.4.3 Non-temporal Properties

The additional examination of aperiodicity, intensity, and skewness indicated that some frequency-domain parameters change along with produced changes in duration. The aperiodicity of both perturbed vowels decreased in production during perturbation. This effect might be a side effect of vowel lengthening, as the longer vowel in the stressed syllable was already less aperiodic in the baseline. Further, less aperiodicity of the perturbed vowel in the Unstressed condition went along with greater intensity of the same vowel, suggesting that the aperiodicity is further coupled with greater intensity. The produced intensity (RMS) increased in the perturbed vowels in both perturbation conditions and in the perturbed fricative in the Stressed condition. We assume the higher intensity to be a consequence of greater emphasis while correcting for the perturbation of the vowels, as seen for other feedback alterations, e.g., delayed auditory feedback (Yates, 1963). In the stressed syllable, greater intensity is not only found for the vowel but also for the fricative. This again calls the μ -gesture model to mind (Saltzman *et al.*, 2008): Word stress gradually spreads its effect from the target of impact (the vowel) on to adjacent segments, with C2 being influenced to a greater degree by lexical stress than C1. The greater intensity along with greater duration also underlines the assumption of Turk and Sawusch (1996) that duration and loudness are processed as a unit. Regarding the interdependencies of the cues with one another (e.g., intensity changes along with compensation to f0), previous research provided quite heterogenous results (see e.g. Patel *et al.*, 2011; and Patel *et al.*, 2015) which could be a matter of linguistic relevance: On the suprasegmental level, prosodic cues might be exchangeable, while on the segmental level, properties such as formant frequencies are unique markers of, e.g., sound quality. This means that alterations of formant frequencies are most likely to be compensated with adjustment of formant frequencies. Further, intensity might indeed be coupled with duration rather than with other parameters, as supported by the current study and studies on delayed auditory feedback.

However, previous studies have concluded that suprasegmental and segmental cues follow common processing mechanisms with the evaluation of local and more global cues (with and without context information, Reinisch *et al.*, 2010; 2011). Duration, in this view, might be anchored in both segmental and suprasegmental levels, which makes a comprehensive attribution to dependencies or independencies with other parameters more complex. Another aspect that shapes the relation of cues is the actual time course of physical events: not all cues are processed at the same time. Spectral cues are used earlier in the perception of vowels than temporal cues (as they are assessable earlier) but are dependent on the context (Reinisch and Sjerps, 2013).

As a general overview of spectral shape, the spectral skewness of the perturbed vowels and fricatives was examined. We found less skewness in the vowels as hypothesized but no effect for the fricatives. Less skewness is the consequence of more energy in the higher frequencies and increased harmonic structure, which might go along with the greater emphasis on the vowel, greater intensity, and less aperiodicity. Saying this, we assume that greater intensity is the actively used cue to emphasize the vowel, which was de-emphasized due to compression in the auditory feedback. However, in examining the relations between the three spectral parameters (intensity (RMS), aperiodicity, and skewness) with each other, only changes in intensity and aperiodicity between hold phase and baseline in the Unstressed condition correlated significantly. Finally, note that we do not regard the changes in non-temporal parameters as specific for stress realization, as all changes in the perturbed vowel occurred in both the stressed and the unstressed syllable.

3.5 Conclusion and Limitations

The current study supports the contention that speakers monitor the surface timing of their own utterances by using auditory feedback information about the timing of the previous and ongoing speech segments. Speakers are flexibly able to adjust segment durations dynamically in the ongoing speech sequence based on the auditory feedback, and can in some cases update the motor control plans accordingly as they unfold. This information is at this time to our knowledge not comprehensively accounted for in current models of speech production, but combines aspects found in the DIVA model (the contribution of auditory feedback to speech planning and execution) and the Articulatory-Phonology/Task-Dynamics framework (timing of gestures as determined by prosodic structure, fur further discussion see Oschkinat and Hoole, 2020; Karlin *et al.*, 2021). The idea of timing mechanisms that are not entirely elaborated on a phonological level (phonology-extrinsic) has also been suggested and discussed recently by Turk and Shattuck Hufnagel 2020. While the perturbation of the unstressed word-initial syllable caused a global lengthening of following segments, the perturbation of the stressed word-medial syllable caused a local compensatory reaction of the syllable's nucleus, accompanied by some adaptive behavior, although only to a small proportion of the online adjustment. The examination of duration on different prosodic levels revealed specific timing strategies that stress the representation of duration as a non-arbitrary property of fluent speech.

Our results underline the specificity of temporal feedback alterations and provide insight into the possibilities for using the temporal perturbation paradigm to further contribute to our understanding of planning and execution of temporal segmental and suprasegmental cues in speech production.

One limitation of our data is that we cannot neatly disentangle the position of the syllable from the stress pattern. The inclusion of both contexts separately would be a fruitful addition to the sparse body of research on speech timing under temporally perturbed auditory feedback – and the small body of research on the influence of prosodic conditions in any form of feedback perturbation. The other limitation of the current study concerns the onset stability and the systematic reaction of C1 and C2 in the face of the perturbation. For a more rigorous conclusion about their temporal behavior, kinematic data is indispensable. The data presents a sample of participants as one group. We observed individual differences in the reaction within that group, with some of the subjects even compensating for the onset CC perturbation. The detailed investigation of individual reaction patterns is beyond the scope of this paper. However, we are keeping the significant amount of variability in mind for future studies, aiming for a better understanding of its nature and its relation to temporal perceptual acuity and non-speech motor variability.

3.6 Appendix A

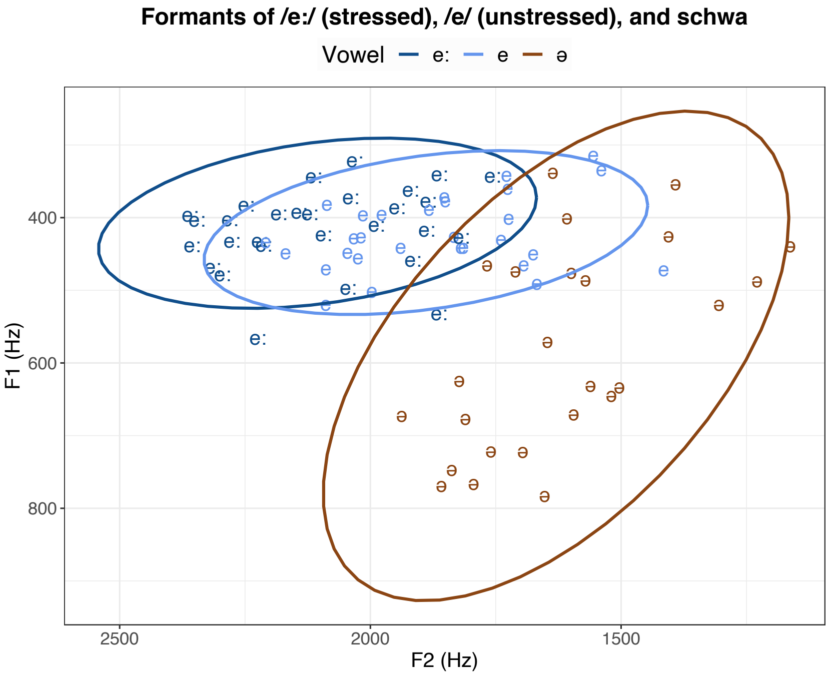


Figure 3.12: First and second Formants (F1/F2) of the three vowels in “Tschetschenen” (/tʃeˈtʃeːnən/). Vowels were provided by the *wrassp* package for signal analysis (Bombien *et al.*, 2021) using *EMuR* (Winkelmann *et al.*, 2020). Formants were extracted over all trials of both perturbation conditions and summarized per vowel per speaker. Formant values were not corrected and should only serve as an overview for typical productions of /e/ in unstressed position, /e:/ in stressed position, and schwa.

3.7 Appendix B

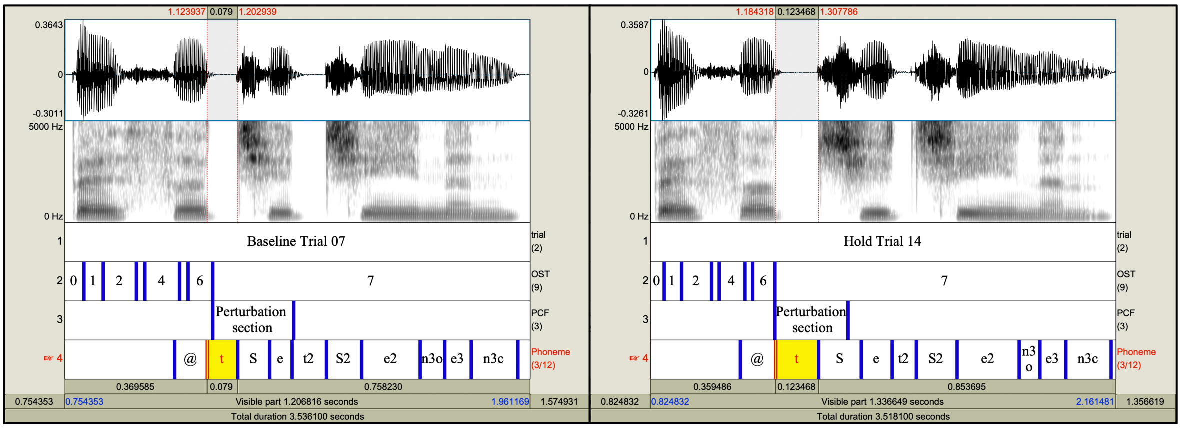

1.123937 0.079 1.202939
0.3643
0
-0.3011
5000 Hz
0 Hz
1 Baseline Trial 07 trial (2)
2 0 1 2 4 6 7 OST (9)
3 Perturbation section PCF (3)
4 @ t S e t2 S2 e2 n3o e3 n3c Phoneme (3/12)
0.369585 0.079 0.758230
0.754353 0.754353 Visible part 1.206816 seconds 1.961169 1.574931
Total duration 3.536100 seconds
1.184318 0.123468 1.307786
0.3587
0
-0.3261
5000 Hz
0 Hz
1 Hold Trial 14 trial (2)
2 0 1 2 4 6 7 OST (9)
3 Perturbation section PCF (3)
4 @ t S e t2 S2 e2 n3 o e3 n3c Phoneme (3/12)
0.359486 0.123468 0.853695
0.824832 0.824832 Visible part 1.336649 seconds 2.161481 1.356619
Total duration 3.518100 seconds

Figure 3.13: Spectrograms with accompanying textgrids as provided by Praat. The second Tier (OST) indicates the different reached stages in the online status tracking, the Tier PCF shows the perturbation section. The example shows a poor fit of the perturbation section in the hold phase (right spectrogram) compared to the baseline fit (left spectrogram). Note that in the baseline trial (left spectrogram) the perturbation section appropriately fits onto the onset and the vowel. In the right spectrogram, the onset consonants [t] and [ʃ] are both much longer than in the baseline trial so that the perturbation section does not cover the vowel anymore (see t durations above the spectrograms in both panels). The second Tier (OST) indicates the different reached stages in the online status tracking.

3.8 Appendix C

The following tables report the significance of the interactions received from the linear mixed models calculated in sections 3.3.1 and 3.3.3. In model 2 (Table 3.4), the threeway-interaction was dropped. “Segment” is the concatenation of sound (e.g, CC) and syllable (e.g., syllable 1).

Table 3.3: Statistical outcome of model 1. CC and V with absolute durations (section 3.3.1).

Type III Analysis of Variance Table with Satterthwaite’s method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
phase	26,680	26,680	1	28.90	35.55	<0.001
Segment	14,649,163	48,883,054	3	9241.87	6505.86	<0.001
condition	36,799	36,799	1	9245.35	49.03	<0.001
phase:Segment	310,077	103,359	3	9241.87	137.71	<0.001
phase:condition	11,198	11,198	1	9244.15	14.92	<0.001
Segment:condition	219,349	73,116	3	9241.87	97.42	<0.001
phase:Segment:condition	197,359	65,786	3	9241.87	87.65	<0.001

Table 3.4: Statistical outcome of model 2. C1, C2, and syll. 3 with absolute durations (section 3.3.1).

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
phase	2,143	2143	1	28.92	1.61	0.215
Segment	73,230,559	18,307,640	4	11570.93	13737.91	<0.001
condition	2	2	1	11575.30	0.00	0.969
phase:Segment	55,301	13,825	4	11570.93	10.37	<0.001
phase:condition	8,666	8,666	1	11573.83	6.50	0.011
Segment:condition	12,811	3,203	4	11570.93	2.40	0.048

Table 3.5: Statistical outcome of model 3. CC and V with relative durations (section 3.3.3).

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
phase	30	30	1	28.85	4.08	0.0528
Segment	181,849	60,616	3	9214.90	8123.19	<0.001
condition	18	18	1	9249.30	2.38	0.1230
phase:Segment	3,702	1,234	3	9241.90	165.37	0.001
phase:condition	57	57	1	9246.96	7.64	0.0057
Segment:condition	2,344	781	3	9241.90	104.69	<0.001
phase:Segment:condition	2,050	683	3	9241.90	91.59	<0.001

Table 3.6: Statistical outcome of model 4. C1, C2, and syll. 3 with relative durations (section 3.3.3).

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
phase	737	737	1	11598.54	78.37	<0.001
Segment	880,950	220,237	4	11596.05	23413.98	<0.001
condition	223	223	1	115596.92	23.67	<0.001
phase:Segment	204	51	4	11596.05	5.43	<0.001
phase:condition	167	167	1	11597.35	17.75	<0.001
Segment:condition	60	15	4	11596.05	1.58	0.18
phase:Segment:condition	209	52	4	11596.05	5.55	<0.001

3.9 Appendix D

Fundamental frequency was estimated like aperiodicity with the Matlab function *yin* (Cheveigné and Kawahara, 2002). The mean f0 values for each vowel segment were entered into the analyses below (in calculating the mean f0, the individual raw f0 values were inversely weighted by the corresponding aperiodicity value to reduce the influence of aberrant f0 values at the vowel margins). The frequency ranges for detecting f0 were set to 150 Hz to 400 Hz for female subjects and 70 Hz to 200 Hz for male subjects.

Fundamental frequency measures were grouped by sex, condition, and phase and all values below or above the 95% confidence intervals were excluded for the following calculations to reduce measurement errors/outliers. Data were transformed to the mel scale for further calculations using the emuR package (v2.1.1, Winkelmann *et al.*, 2020).

The majority of subjects produced the first (unstressed) syllable with a higher pitch than the second stressed syllable. The left panel of Figure 3.14 shows f0 values in mel for the perturbed vowel in both conditions in the baseline (green box), hold phase (yellow box), and aftereffect phase (magenta box) of the experiment split by sex. The right panel additionally shows differences from the baseline mean in semitones for both sexes together for a more accessible overview in perceptual terms.

A linear mixed model was fit with f0 in mel as dependent variable and phase, condition, and sex as predictors with an interaction between phase and condition. The interaction between phase and condition was added as a within-subject random effect (intercept and slope). *Emmeans'* pairwise comparison between the two phases per condition revealed no significant difference for either condition (Unstressed condition: H-B estimate = 2.79; SE = 1.79; df = 29; *t-ratio* =1.563; *p* = 0.129; Stressed condition: H-B estimate = -3.07; SE = 1.91; df = 29; *t-ratio* = -1.607; *p* = 0.119).

Note here that these results need to be interpreted with caution, since the analysis does not take into account different realized intonation patterns of the stress pattern.

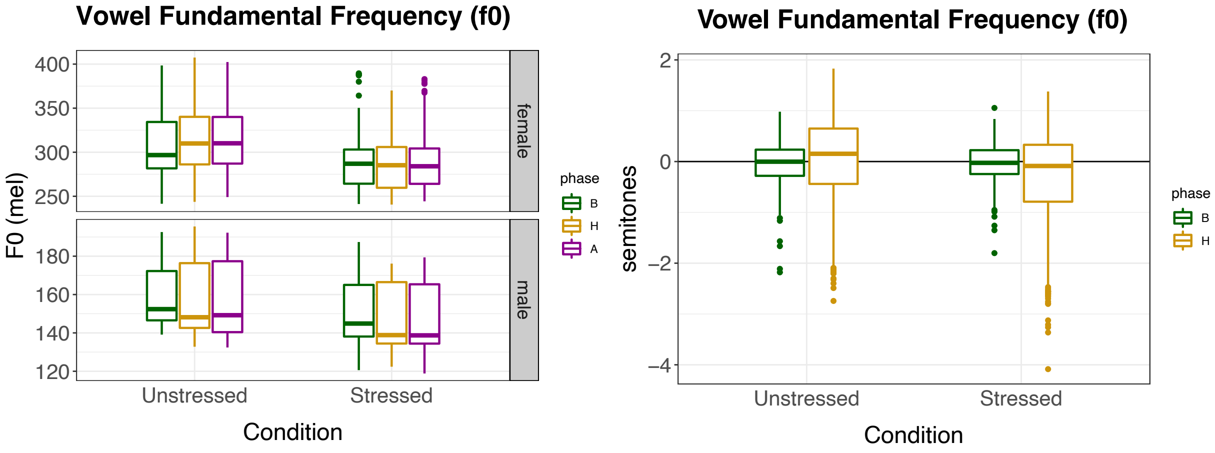


Figure 3.14: The left plot shows f0 values for all subjects in mel on the y-axis and condition on the x-axis split by Sex. Different phases of the experiment are marked with different colors (Baseline: Green, Hold: yellow, Aftereffects: magenta). The right panel shows productions in semitones relative to the baseline mean per condition.

Chapter 4
Compensation to temporal Auditory Feedback Perturbation and its Relation to general Motor Stability and Auditory Acuity

Abstract

Spectral auditory feedback perturbations indicated a link between feedback and feedforward mechanisms in speech production when subjects compensated for applied shifts. Thereby, it was shown that subjects with a higher perceptual auditory acuity compensate more (Villacorta et al., 2007). However, the reaction to feedback perturbation is not merely a matter of perceptual acuity but is also certainly affected by predicting and producing precise motor action. This interplay between prediction, perception, and motor execution seems to be crucial for the timing of speech and non-speech motor actions. The present study links responses to *temporally* perturbed auditory feedback to general rhythmic abilities in production and perception. We expect better auditory acuity to be connected with more compensation as found for spectral perturbations. Further, we expect greater variability in general motor timing tasks linked to more compensation since a less stable motor system should also be more malleable in perturbation. Both auditory acuity and motor stability were shown to affect responses to temporal auditory feedback perturbation, but with a greater weighting of one over the other dependent on the prosodic structure of the perturbed sequence.

4.1 Introduction

The interaction of feedback and feedforward systems in speech production has been a major focus of interest in speech research. Real-time auditory feedback perturbations investigated the influence of auditory feedback on speech production in moment-to-moment control and for planning future productions. While the contribution of auditory feedback to planning and controlling spectral properties of speech sounds has been thoroughly studied, beyond the investigations in chapters 2 and 3, only a few studies investigated the role of auditory feedback for speech timing (see, e.g., Cai *et al.*, 2011; Mitsuya *et al.*, 2014; Floegel *et al.*, 2020). The studies presented in chapters 2 and 3 showed that auditory feedback is indeed used to control and plan temporal properties of speech, but with syllable structure (chapter 2) and stress pattern and syllable position (chapter 3) shaping the responses. The experiments pointed out that for comprehensively modeling speech production, two aspects need to be combined: the incorporation of auditory feedback on planning and control level (as in the DIVA model, Guenther, 2006) as well as prosodically induced timing relations (as elaborated in Articulatory Phonology/Task-Dynamics, Browman and Goldstein, 1992).

In studying the relevance of auditory feedback for speech production, previous investigations have focused on whether individual differences in the preciseness of discriminating auditory information (*auditory acuity*) relate to individual differences in speech production. Findings from these investigations mainly were in line with assumptions elaborated in the DIVA model. The DIVA model assumes speakers to have spatio-temporal target regions for speech segments. Those targets are established via auditory and somatosensory feedback in speech acquisition (Guenther, 2016, p.131). The size of the acquired speech targets is assumed to depend on auditory acuity and sensory error detection performance. Thereby, speakers with better auditory acuity establish smaller speech targets, resulting in more distinct productions of different speech sounds and less variability in production. Accordingly, speakers with poorer auditory acuity establish larger speech targets causing less distinct productions of different speech sounds and higher variability (Perkell *et al.*, 2004a; Perkell *et al.*, 2004b; Perkell *et al.*, 2008; Ghosh *et al.*, 2010). Individual differences in auditory acuity became a further focus of interest in connection with auditory feedback perturbation studies. Villacorta *et al.* (2007) assessed auditory acuity for F1 discrimination in vowels and set them in relation to reactions to spectral upwards and downwards alterations of F1 in the same vowels. They found that the better the individual auditory acuity, the more the speaker compensated for the applied feedback alteration. This conclusion was also drawn by Brunner *et al.* (2011) for perturbed consonants. They found speakers with a higher auditory acuity to produce /s/ and /ʃ/ with a more distinct acoustical contrast and to use compensation strategies to a greater extent than low acuity speakers.

While the role of auditory feedback for speech production is examinable with auditory feedback alterations, it is more difficult to assess the contribution of somatosensory feedback for reaching speech targets. In speech production, however, speakers rely on both auditory and somatosensory feedback. Ghosh *et al.* (2010) inquired into the relation of produced sibilant contrast to auditory and somatosensory acuity. Positive correlations indicated that better auditory and better somatosensory acuity are connected with a more distinct acoustic sibilant contrast. The absence of a relation between the two sensory modalities suggested that both auditory and somatosensory feedback contribute to the production of sibilant contrast but independently.

Similar to the two interacting sensory feedback pathways, speech production relies on both sensory feedback and feedforward mechanisms. While individual abilities in feedback control have been considered crucially influencing factors in building and controlling speech targets, much less attention has been given to the thought that also feedforward mechanisms, more precisely motor execution abilities, are governed by limits of individual abilities. In spectral auditory feedback perturbation, a better auditory acuity was shown to lead to more compensation. Another aspect that is likely to influence distinctiveness in speech production is the ability to execute motor commands for desired speech targets precisely. However, not much research has addressed the latter. Our study in chapter 2 pointed out that, at least for timing relations in speech, different prosodic structures cause segments to be more stable or less malleable in their articulatory execution than others. With a novel version of Audapter (Cai *et al.*, 2011; Tourville *et al.*, 2013), we focally manipulated absolute durations of onset, nucleus, and coda in a similar phonological context. Results showed that speakers compensate in the temporal domain as they do in the spectral domain, but with dependence on syllable structure: While speakers compensated and adapted for durations of nucleus and coda, they did not compensate or adapt for alterations of onsets in onset + vowel perturbation. The data indicated that auditory feedback is used to control timing mechanisms in speech and that the stability of gestural coordination conditions articulatory adjustments. The more stable coupling relation of onsets is more resistant to adapting to errors introduced by auditory feedback. Another possible explanation was given by assuming that for speech timing, somatosensory feedback might be used more in onsets for correction mechanisms, as the auditory feedback is always delayed and cannot give information in onsets about timing relation in larger prosodic units such as the syllable or the word. The study conducted in chapter 3 replicated the findings for onset stability.

Both of our investigations in chapters 2 and 3 suggested differences in structural motor stability on the feedforward side. These structural differences further lead to the question of whether *individua*l differences in motor execution also shape timing mechanisms. The contribution of individual motor stability and individual auditory acuity to timing mechanisms in speech production will be the substance of the following study.

One approach towards testing general motor execution abilities is provided by rhythmic finger tapping, which inherently carries a timing component. In many scientific tapping paradigms, subjects tap with the index finger of their writing hand at an individual regular tempo for a period of time. Dalla Bella *et al.* (2017) built a battery of perception and finger tapping tasks for assessing complex individual timing profiles for persons with and without timing disorders. Typical tapping tasks comprise unpaced tapping tasks, where the subject is meant to tap at an individual rate without a guiding beat and paced tapping tasks, where the subject is supposed to hit an accompanying beat or synchronize to music or speech (Dalla Bella *et al.*, 2017). Unpaced tapping tasks give the examiner insight into feedforward timing mechanisms and their stability in motor execution (see Drake *et al.*, 2000). Tapping to a beat, on the other hand, tests for sensorimotor synchronization (see Repp and Su, 2013 for an overview).

A link between general rhythmic executive abilities and speech production was found in the examination of tapping performance in healthy speakers and speakers with pathologic speech timing disorders. Falk *et al.* (2015) found weaker synchronization abilities with a metronome or a musical stimulus in children and adolescents who stutter than in nonstuttering peers. Thereby, individuals who stutter showed weaker performance regarding consistency and accuracy in tapping, the latter induced by over-anticipation of the pacing event.

The connection between non-speech timing abilities and rhythmic deficits in speech production suggests similar timing mechanisms underpinning speech and non-speech domains (Nozaradan *et al.*, 2012; Peelle and Davis, 2012). This conclusion was further specified for timing mechanisms concerning forward prediction in music and speech (Iversen *et al.*, 2009; Tierney and Kraus, 2014). Further research also found that in the coordination of speech and non-speech motor action, mechanisms are coupled in magnitude and temporal coordination with multisensory feedback. For example, Gentilucci *et al.* (2004) measured the opening gesture of the vowel in the syllable /ba/ and found speakers to open the jaw more when they moved an apple to their mouth than a cherry. In temporal coordination, subjects coordinated the apex of the pointing gesture with the stressed syllable of the uttered pointing item (Rochet-Capellan *et al.*, 2008). Further, the coproduction of speech with finger tapping indicated that emphasis in one domain (e.g., stressing a syllable) affects the other domain as well (e.g., more emphasized tapping, Parrell *et al.*, 2014). All these studies indicate a link between motor actions of different domains. The link between speech and non-speech motor action is of significant interest when investigating the role of feedforward stability for timing mechanisms in fluent speech. Therefore, individual differences in the stability of temporal speech representations might be predicted by individual non-speech motor behavior.

The current study probes the link between individual abilities in paced and unpaced finger tapping tasks, auditory acuity, and reactions to temporal auditory feedback alterations. Thereby, this chapter aims to shed light on the influence of individual auditory acuity and general feedforward stability on speech production. With this aim, we address both feedback and feedforward systems as key actors for successful speech production.

We expect speakers with a higher auditory duration discrimination ability (auditory acuity) to compensate more for temporal feedback alterations as found for spectral properties of speech. Moreover, we expect speakers with a worse performance in motor execution in finger tapping tasks to compensate more. This hypothesis ties up to the findings in chapters 2 and 3, where a less stable system was more malleable in the face of a temporal perturbation. The investigation is essentially exploratory in nature and is likely to cause complex outcomes, which will be interpreted and discussed with a primary interest in giving future research guidance rather than predicting compensatory behavior.

4.2 Methods (Procedure and Data Processing)

The following sections outline the three testing blocks Perturbation, Tapping, and Perception. Forty-five native speakers of German performed all three testing blocks in one testing session of approximately 2.5 h. Participants were between 19 and 30 years of age (mean age: 23y, 34 females) and received financial compensation for their participation. None of the subjects claimed to have any speech, voice, or hearing disorders. All of the subjects started with the auditory feedback perturbation block. After that, the order of blocks two and three was counterbalanced over subjects. Blocks two and three consisted of tapping and perception tasks from the *Battery for the Assessment of Auditory Sensorimotor and Timing Abilities* (BAASTA, Dalla Bella *et al.*, 2017) extended by tasks incorporating speech stimuli outlined below.

4.2.1 Temporal auditory Feedback Perturbation

The temporal auditory feedback experiment tested the sensitivity to temporal perturbations in onset, vowel and coda of a syllable. The data was taken from the perturbation experiment in chapter 2 (Oschkinat and Hoole, 2020), including the same 34 subjects in the Onset condition and 33 subjects in the Coda condition after scanning the data for correct triggering of the perturbation section. In chapter 2, the procedure and methods have been extensively explained and will therefore not be outlined again in this section to avoid tiring the reader with repetitions. The following section briefly summarizes the measures of interest. If one desires to refresh the procedure, we would kindly point the reader to chapter 2, sections 2.2 and 2.3.

The perturbation data analyses in chapter 2 were performed with word-normalized durations relative to the baseline mean in the Hold phase relative to the baseline for each segment of interest (CC /pf/ and V /a/) per perturbation condition (Onset and Coda condition). Accordingly, four compensation measures are considered in the following calculations: Compensation to the onset segment in the Onset condition (Onset CC), compensation to the vowel in the Onset condition (Onset V), compensation to the vowel in the Coda condition (Coda V), and compensation to the coda segment in the Coda condition (Coda CC). These measures are presented visually in Figure 4.1, which is a repost of Figure 2.4 from chapter 2. Note that the Onset CC and the Coda V were stretched in perturbation so that a compensatory reaction is indicated by a shortening of productions (negative estimates relative to the baseline mean). For the following measures, the values of the onset CC compensation and the coda V compensation were multiplied by -1. Thus, a compensatory response is always indicated by a positive value and following the perturbation direction by a negative value.

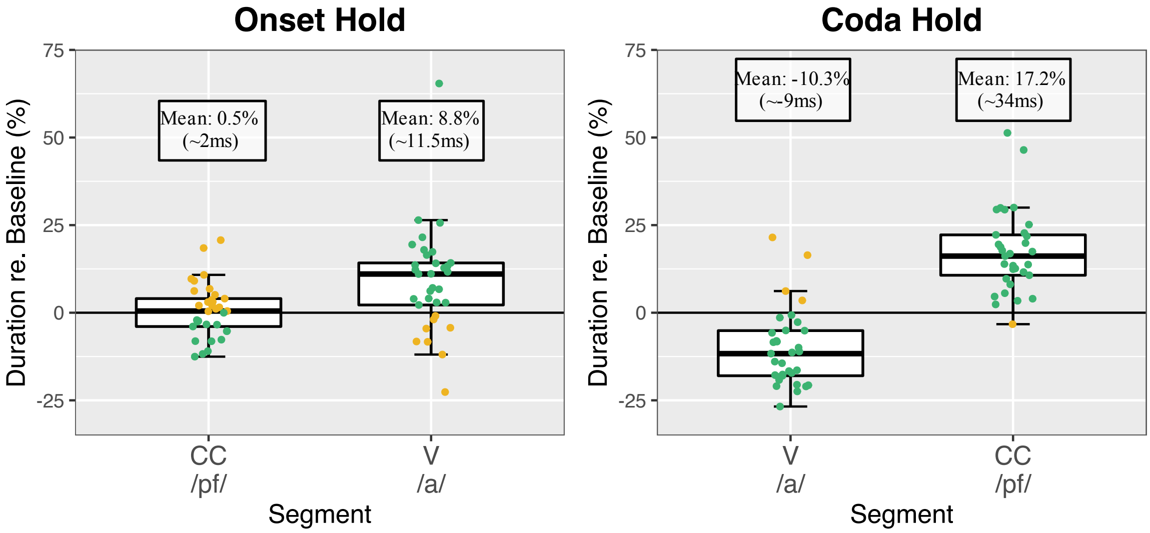


Figure 4.1: Repost of Figure 2.4 from chapter 2. Boxes present the four compensation measures of interest (normalized relative durations in the hold phase relative to the baseline mean, Onset CC and Onset V in the left panel (34 subjects), Coda V and Coda CC in the right panel (33 subjects)). Individual subjects are represented with colored dots where green dots mark the compensatory behavior and golden dots mark a following of the perturbation direction. 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.

4.2.2 *Tapping Battery*

For the tapping test block, subjects were seated in front of a Roland SPD-6 MIDI percussion pad linked via a Midi-Interface (Miditech, midiface, 4x4) to a computer controlled by MAX-MSP software (version 6.0). Loudspeakers delivered sound stimuli with a fixed volume which was kept constant over subjects. The experimenter instructed the subject to tap with their writing hand's index finger and directed the procedure. Practice trials preceded each of the following tasks, which could optionally be skipped when the following task was very similar to the preceding one. Tasks one, two, and six are adopted from the BAASTA framework (Dalla Bella *et al.*, 2017). Tasks three, four, and five contain speech stimuli of different complexity implemented for this study's particular purposes.

1) Unpaced spontaneous tapping

In the first task, subjects tapped at a self-chosen comfortable tapping rate for 60 seconds to assess their motor variability. Participants were instructed to tap as regularly as possible.

2) Paced metronome tapping

The second task tested for synchronization ability when an isochronous sequence of metronome beats (tone frequency: 1319 Hz) was auditorily presented for a period of 60 beats. Participants were required to tap to the perceived beat in three runs, whereby the inter-onset-intervals (IOI in ms) of the beats differed (first run: 600 IOI, second run: 750 IOI, third run: 900 IOI).

3) Syllable tapping

The third task was similar to task 2), except that synchronization to a simple speech element was tested. The presented stimulus consisted of the syllable "*bla*" spoken by a female speaker. That same speaker produced all following speech task stimuli. The spoken syllable was spliced together at psychoacoustically isochronous intervals based on the p-center algorithm from Cummins and Port (1998). Three tempi were performed in three consecutive runs (IOI of 600 ms, 750 ms, and 900 ms).

4) Wordlist tapping

The second speech task required tapping to a spoken wordlist of real monosyllabic words with complex onsets (CCV(C)) to test for synchronization to more complex speech material. The words were temporally arranged to realize a psychoacoustically isochronous sequence for human perception based on the p-center algorithm from Cummins and Port (1998) mentioned above. In some cases, corrections were made, e.g., if there were several p-centers per word, the one closer to the vowel was chosen, and if the algorithm did not detect a p-center, the second zero crossing of the vowel served as a reference. An IOI of 600 ms was chosen in view of previous studies that classified this tempo as a natural medium rate in human perception and production, close to the heartbeat.

Wordlist: *klein; blond; klar; klug; schlau; schlecht; schlimm; stark; stumm; still; blau; blöd; Kleid; Blatt; Block; Blitz; Klo; Klang; Staub; Sturm; Stich; Schlamm; Schlauch; Schleim*

5) Sentence tapping

The third of the speech synchronization tasks included fluent speech in the form of rhythmic spoken sentences. Unlike all previous stimuli, the sentence task did not provide moments of silence between the beats (or between the sentences) but rather a continuous sound flow. The trochaic rhythm of the sentences suggested tapping on every second syllable of the phrase. The inter-tap interval between two syllables with an accompanying beat was 600 ms.

Sentence (tap target syllables in bold)*: Der **Va**ter **fährt** den **LKW**, die **Mö**bel **trägt** der **Sohn** ins **Haus**. Aufs **So**fa **steigt** der **Groß**pa**pa**, die **O**ma **sorgt** sich **um** sein **Wohl**. Die **Frau** gibt **nun** dem **Herrn** den **Brief**, die **Da**me **reicht** dem **Herrn** das **Tuch**.*

6) Music tapping

The sixth task tested synchronization to music. Two piano midi stimuli were created from well-formed (regular) excerpts of the beginning of Bach's *Badinerie* and Rossini's *Wilhelm Tell*. Both excerpts were set to an inter-quarter-interval of 600 ms and presented for a duration of 64 beats. Rossini qualified *Wilhelm Tell* for the more challenging of the two pieces as he composed it starting with an upbeat.

All tapping data were pre-processed to extract three main qualities of the tapping performance: motor variability, synchronization consistency, and synchronization accuracy, as previously done by Dalla Bella *et al.* (2017). For tasks 1 to 4, the first ten taps were discarded, and artifacts (inter-tap intervals below 100 ms) and outliers were removed. For all tasks, including the unpaced tapping task, the mean inter-tap-interval (ITI) was calculated, and the coefficient of variation of the ITI (CV of the ITI, namely, the ratio of the SD of the ITIs over the mean ITI) was taken as a measure for *motor variability.* For the paced tapping tasks, circular statistical analyses were performed to classify synchronization performance (for more information on the advantages of linear and circular analyses, see Dalla Bella *et al.*, 2017). In the circular statistics framework, taps were presented in a 360-degree polar scale, where the pacing event or p-center in the speech stimuli is set at 0. The tap is positioned as an angle relative to the pacing stimulus. The taps' angles were treated as unit vectors and were used to calculate the mean resultant vector R (Fisher, 1995; Berens, 2009; Mardia and Jupp, 2009; Dalla Bella *et al.*, 2017). With the vector R, synchronization consistency (i.e., the reciprocal of variability) and synchronization accuracy were calculated (Sowiński and Dalla Bella, 2013; Dalla Bella *et al.*, 2017). The vector length indicates *synchronization consistency* (ranging from 0 to 1) and reflects the variability of the time lag between the taps and the pacing stimuli. For example, if the taps always occur exactly 200 ms later than the pacing stimulus, vector length would be 1 (indicating perfect consistency).

Synchronization accuracy was obtained by measuring the angle of vector R (relative phase, in degrees). The angle indicates whether the tap was placed before (negative angle) or after (positive angle) the pacing stimulus. Accuracy was only computed if the synchronization performance was above chance (Falk *et al.*, 2015; Dalla Bella *et al.*, 2017). Accuracy values were log-transformed for further calculations.

4.2.3 Perception Tasks

The third block tested for individual perceptual abilities. Five adaptive staircase tasks assessed individual auditory acuity performances for temporal properties of various stimuli types. The listener was seated in front of a computer and provided with headphones. Volume was set to a comfortable level as tested and determined by the experimenters and was not changed between listeners unless requested. After the experimenter started the procedure in MATLAB, listeners performed the tasks by entering their responses directly into the testing computer. The first three staircase tasks captured duration discrimination abilities in a 2-interval 2-alternatives forced-choice paradigm which required judgments of the two perceived stimuli as identical or different.

In addition to the three discrimination tasks, two beat-alignment tasks related to the sentence and music tasks (5 and 6) of the tapping battery in section 4.2.2 were performed. Task four and five required beat-alignment judgments in a 1-interval 2-alternatives forced-choice paradigm. The decision required a binary judgment on whether the accompanying metronome beat perfectly aligned with the presented speech or music stimulus or not.

For all five tasks, continua of stimuli between two endpoints were generated in praat, where one endpoint consisted of the original stimulus and the second endpoint of a manipulated version. The manipulations included duration differences exclusively. For the three duration discrimination tasks (1-3), stimuli were presented in pairs (2-interval) in which one stimulus was always the original stimulus and the other stimulus varied in degree of manipulation between the two endpoints. Manipulations of segment durations were performed in praat using psola (see task-specific description below). The presented stimuli were randomized, whereby the original stimulus was either in the first or in the second position. In the two beat-alignment tasks (4 and 5), one endpoint of the continuum was a stimulus with perfect beat-alignment, and the other endpoint a stimulus with the maximally shifted beat. In beat-alignment tasks, always one stimulus from the continuum was presented while the degree of alignment shift varied along the continuum. The difference between the two presented stimuli in tasks 1 to 3, or the degree of metronome shift in tasks 4 and 5 is referred to as *delta* (in ms). In each task, the delta could be varied in increments of 1 ms. Estimations of the reached delta (based on calculations described further below) will serve to measure each subject's individual *auditory acuity*.

1) Pure tone duration discrimination

The first task comprised the presentation of two pure tones (frequency: 333.3 Hz) that differed in duration exclusively. The stimulus at one endpoint had a duration of 600 ms, and the other (manipulated) endpoint stimulus was 1200 ms long. Hence, the maximum difference between the two presented stimuli was 600 ms. The pair with maximum difference served as start delta (first presented stimulus pair), and a total of 600 pairs of stimuli were provided for presentation.

2) Onset duration discrimination

In the second task, a monosyllabic CVC word ("Schaf", /ʃaːf/, sheep) was provided as one endpoint of the continuum. For the other endpoint, the onset consonant and the vowel were temporally altered analogously to the auditory feedback perturbation paradigm's Onset condition (see chapter 2).

Accordingly, the onset /ʃ/ was stretched by 200 ms while the following vowel /aː/ was compressed by 200 ms compared to the original stimulus, resulting in a start delta of 200 ms. For presentation, 200 pairs of stimuli were provided.

3) Coda duration discrimination

Another monosyllabic CVC word ("Gas", /gaːs/, gas) was provided as one endpoint of the continuum in the third task. For the other endpoint, the onset consonant and the vowel were temporally altered analogously to the auditory feedback perturbation paradigm's Coda condition (see chapter 2). Accordingly, for the other endpoint, the vowel /a/ was stretched by 150 ms and the coda /s/ compressed by 150 ms, resulting in a start delta of 150 ms. For presentation, 150 pairs of stimuli were provided.

4) Speech beat-alignment

The speech beat-alignment task included a trochaic sentence with a metronome beat on every stressed syllable in a stable tempo of 600 ms inter-beat-interval. The sentence was spliced together 3 times to create a longer continuous stimulus. Beat-alignment with the speech stimulus followed the p-center algorithm from Cummins and Port (1998) as described before. Misplacement of the metronome was implemented by shifting the beat several milliseconds to the right (later than the original p-center position). The metronome's maximum displacement was 200 ms later than the metronome in the initial perfect beat-alignment stimulus. Hence, the second endpoint of the continuum differed by 200 ms from the original one. Accordingly, the start delta was 200 ms, and 200 different stimuli pairs were provided for presentation. The metronome started after four beats in the sentence (on "zahm") to allow for an initial prediction without the beat.

Sentence (stressed syllables in bold): Der **Bi**ber **ruht** im **war**men **Bau**, der **zah**me **Braun**bär **wohnt** im **Zoo**, das **klei**ne **Mäus**lein **fiept** im **Rohr**.

5) Music beat-alignment

The music beat-alignment task consisted of a midi excerpt of Bach's Badinerie (as in the BAASTA tapping battery, see section 4.2.2). Maximum manipulation resulted in an endpoint stimulus of a 200 ms delayed beat, creating the start delta of 200 ms and a set of 200 presentable stimuli pairs. The beat started four beats after the music started.

All five staircase tasks had fixed step intervals that were adaptive, meaning triggered by the listener's response. Following a mismatch identification, the current delta was multiplied by 0.5; with every not detected mismatch, the delta was multiplied by 1.5 (see Figure 4.2). The tasks always required two correct difference detections in a row to the same stimuli pair to mark a successful identification, whereby the two stimuli in both presented trials appeared in random order. Whenever there was a change of response quality (difference detected/difference not detected), one *reversal* was counted (see Figure 4.2). Each task ended when a fixed number of reversals was reached (12 reversals for the discrimination tasks (1-3), eight reversals for the beat-alignment tasks (4 and 5)).

With this 2-alternative forced-choice paradigm, it remains the constraint that listeners could always claim that they heard a difference between the two presented stimuli or a not well- aligned beat (even if they did not), which would result in a perfect auditory acuity score. For this instance, so-called *catch-trials* were included in each test. Catch-trials are presentations of two identical stimuli or a perfectly aligned beat (delta = 0) to provoke the answer that there was no difference between the stimuli. Catch-trials ensured that subjects actively participated in the task rather than randomly responding and ensured that they were auditorily capable of performing the task. Catch-trials were presented approximately every fourth stimulus, resulting in 4 to 6 catch-trials per task.

Listeners who did not identify more than 50% of the presented catch-trials correctly or did not reach a score below 70% of the start delta of a test were suspected of answering by chance or classified as incapable of performing the task. Hence, they were excluded from that test for further calculations.

To measure individual performance, for every subject and task, an individual auditory acuity score was assessed by calculating a mean over the delta of the most stable six reversals in each task (the six reversals with the lowest SD, indicating a stable pattern of response), visualized in Figure 4.2. The most stable six reversals per test were chosen following Brunner *et al.* (2011). Although in Brunner *et al.* (2011), the most stable 15 trials were chosen, their subjects performed four runs with 80 trials/14 reversals for one auditory acuity test. Since we had much shorter staircase paradigms due to the large test battery and a different paradigm, we decided to include the most stable sequence of 6 reversals. In some cases, the sequence comprised even more than 15 trials per sequence (see, e.g., Figure 4.2), but recall here that every trial had to be identified correctly twice, which leads to more or less stimuli more rapidly when the number of reversals changes.

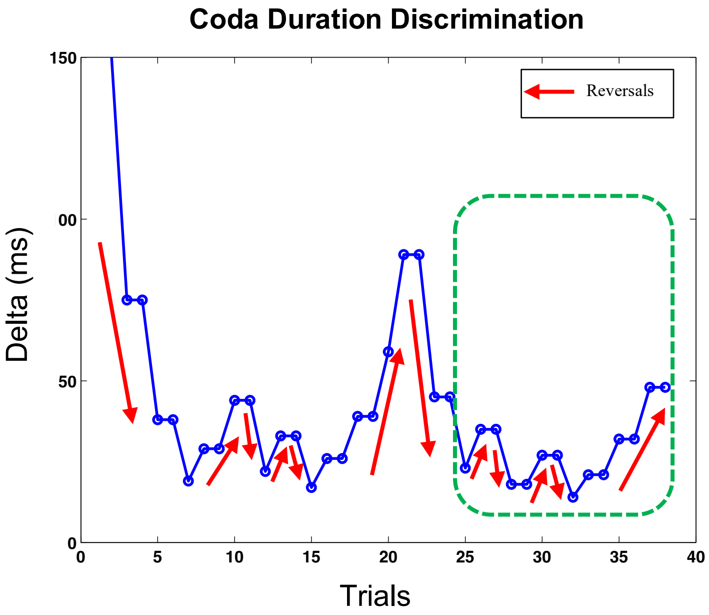


Figure 4.2: Visualization of the course of a perception staircase test (here: coda duration discrimination task). Blue circles indicate one presented stimuli pair. If the listener identified a pair correctly, the same pair was presented again, whereby the order of the two stimuli was randomized. After responding to the second presentation correctly, the delta of the following stimuli pair dropped; with every (single) wrong answer, the delta increased. The test ended after 12 reversals; each reversal marked with a red arrow. The green dashed square indicates the six reversal points (counting on from the reversal at trial 25) with the lowest SD; the mean delta of these trials served as an auditory acuity score for further calculations. Catch-trials are not included in the figure.

4.3 Analyses

The data provided by the tapping and perception test blocks capture many aspects of motor and perceptual abilities concerning speech and non-speech tasks. Per block, subjects performed different tasks, and in the tapping block, three different measures per task were extracted. Since this examination yields a vast dataset, principal component analyses for the tapping and perception block were conducted to avoid correlating every single task parameter with every other.

4.3.1 Principal Component Analysis

For the following calculations, the perception and tapping data were submitted to principal component analyses (PCAs) using R's *mclust* package (Scrucca *et al.*, 2016). The PCA reduces the number of independent variables to single components by extracting the underlying dimension for variables that highly correlate with each other. The extracted underlying dimensions (principal components) of a PCA do not correlate with each other and describe the dataset's maximum variance. In the following, the main components are extracted and used for further calculations.

Before calculating PCA, the data was partitioned. That is, one could throw all data from all tasks into the PCA. This approach, however, could lead to non-interpretable underlying dimensions and might not enable distinguishing perceptual from motor abilities or different qualities of tapping performance from each other. Therefore, PCAs were calculated over all tasks of the perception block and for all tasks *per obtained measure* of the tapping block. Hence, one PCA was performed for perception, one PCA for tapping motor variability, one PCA for tapping synchronization accuracy, and one PCA for tapping synchronization consistency. With this data partitioning, we hoped to keep one general underlying dimension per measure of interest and further expected the PCA to give more insight into the characteristics of the single tasks.

4.3.1.1 Data pre-processing

As PCAs cannot deal with incomplete data, the data was scanned for missing values. Per tapping measure (motor variability, synchronization accuracy, synchronization consistency), subjects with no data in more than three out of the tapping tasks were not submitted to the respective PCA. Four subjects were excluded from further calculations based on this criterium for all three measures. For those subjects who had missing data in three or fewer tasks, NAs were filled with the *k-nearest-neighbor imputation* (knn-Imputation) method (Beretta and Santaniello, 2016). For motor variability, one to three missing values were filled for five subjects. For tapping consistency, three subjects had one to three missing values filled with knn-imputation, and for tapping accuracy, four subjects had one to three filled missing values.

The unpaced tapping task differed from the other tapping tasks in modality, as it was the only task without a pacing event. It gives insight into pure feedforward stability without a guiding stimulus. Therefore, the motor variability of the unpaced tapping task was individually observed in addition to the principal components (and therefore not submitted to the PCA for motor variability).

Before submitting the items (single tasks) to the PCAs, the Kaiser-Meyer-Olkin measure (Kaiser, 1970) verified the measure sampling adequacy (MSA) overall per measure block and single task. An MSA value above 0.5 qualified the single measures for submitting them to the PCA, and the overall MSA measure classified the whole task block as suited for PCA if the overall MSA was > 0.5. For the perception tasks, overall MSA was 0.58. In tapping motor variability, overall MSA was 0.7, whereby the single MSAs for the Rossini music tapping task and wordlist tapping were < 0.5 and hence not submitted to the PCA. In tapping consistency, overall MSA was 0.73, whereby tapping to the sentence was < 0.5 and dropped for further calculations, and the overall MSA for tapping accuracy was adequately 0.72.

Values were centered and scaled when submitted to the PCA. The PCs that explained the most variance defined by the Kaiser criterion (variance > 1; Kaiser and Dickman, 1959) were kept for further calculations. Those comprised the first two principal components per PCA. Figure 4.3 visualizes the first two components for each of the measures of interest.

4.3.1.2 *Interpreting PC scores*

Tables 4.1 and 4.2 summarize the factor loadings of PC1 and PC2 per PCA for each of the single tasks. For the tapping PCAs (Table 4.1), the metronome and syllable tapping tasks show strong loadings on PC1, while the music tapping tasks show strong loadings on PC2. Tapping to the sentence seems to be clustering on PC1 and PC2 equally strongly (or weakly). PC1 is therefore interpreted as a general measure for motor variability, accuracy, or consistency. Adding to that general measure, PC2 reflects musical abilities, or the ability to find rhythm in fluent/continuous sound stimuli (sentence and music tapping) vs. non-continuous stimuli (metronome, syllables, and wordlist tapping). For all tapping measures, *better performances*, meaning low motor variability values, low (close to zero) accuracy values, and high consistency values, are associated with *lower* PC1-scores.

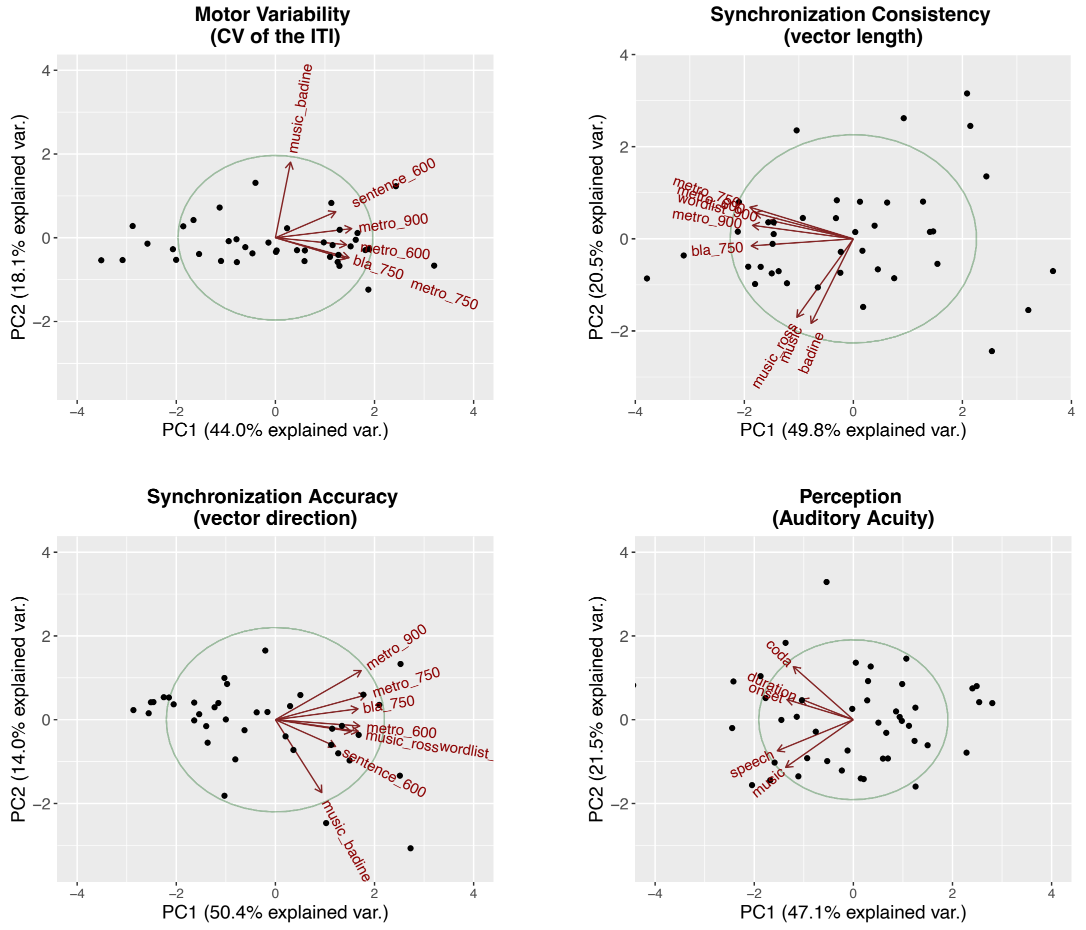


Figure 4.3: Visualization of the first (x-axis) and the second (y-axis) principal components along with their amount of explained variance per PCA. Dots mark the single subjects; vectors represent the factor loadings of the single tasks on each of the components. Tasks are abbreviated by type (e.g., "metro" for metronome tapping) followed by the IOI (if relevant).

For the perception PCA (Table 4.2), all of the tasks correlate negatively with PC1, with a *higher* PC score indicating a better perception (a lower auditory acuity threshold). Therefore, PC1 reflects general auditory acuity. The music beat-alignment perception task correlates highly with PC2, as does the speech beat-alignment task, although less intensely. This clustering indicates that PC2 reflects musicality and beat-alignment judgments. This interpretation is further encouraged by the coda perception task, which correlates the other way round with PC2. In finding the rhythm in speech, the p-center serves as a landmark for temporal orientation in synchronic rhythmic speech perception. While the p-center is located in the transitions between the onset and the vowel of a syllable, an excellent perceptual discrimination ability in codas might not improve beat-alignment with speech (aligning the beat with the onset of syllables).

Table 4.1: Factor loadings for each of the tapping tasks on the PCs for the three PCAs (Motor Variability, Consistency, Accuracy). High factor loadings on a component (> 0.3) are shaded in grey. MV stands for Motor Variability.

	PC1 MV	PC2 MV	PC1 Consistency	PC2 Consistency	PC1 Accuracy	PC2 Accuracy
metro_600	0.45	-0.08	-0.43	0.21	0.39	-0.06
metro_750	0.44	-0.24	-0.45	0.25	0.41	0.26
music_badine	0.10	0.88	-	-	0.21	-0.75
music_ross	-	-	-0.25	-0.63	0.35	-0.12
metro_900	0.48	0.11	-0.44	0.11	0.39	0.51
bla_750	0.47	-0.23	-0.44	-0.06	0.38	0.11
sentence_600	0.38	0.30	-	-	0.27	-0.27
wordlist_900	-	-	-0.36	0.15	0.38	-0.12

Table 4.2: Factor loadings for each of the perception tasks on the PCs for the Perception PCA. High factor loadings on a component (> 0.3) are shaded in grey.

	PC1 Perception	PC2 Perception
onset	-0.46	0.24
coda	-0.41	0.64
duration	-0.36	0.25
speech	-0.52	-0.38
music	-0.47	-0.57

4.3.2 *Data Summary*

The following calculations aim at examining the contribution of general motor abilities (tapping) and perceptual abilities (perception) to the responses to temporal auditory feedback perturbation. For this purpose, we develop linear multiple regression models to understand better how the single predictors relate to compensation. The single predictors, along with the single compensation measurements incorporated into the following analyses, are summarized in Table 4.3. We would like to note once more at this point that the analyses are exploratory. Focal real-time temporal auditory feedback perturbation is a very recent method, and not much is known about how speakers control and update temporal properties of speech and patterns of timing in speech production via auditory feedback. Consequently, even less information is available about what mechanisms might shape these responses. The extensive body of different perception and tapping tasks aims to paint a wide-ranging picture of possible influencing factors.

Before further calculations, outliers of the generated first and second principal components and of the spontaneous tapping task (data outside the 95% confidence intervals) were set to Na (PC1 Motor Variability: 0 subjects, PC2 Motor Variability: 3 subjects, spontaneous Tapping Motor Variability: 2 subjects, PC1 Consistency: 0 subjects, PC2 Consistency: 4 subjects, PC1 Accuracy: 1 subject, PC2 Accuracy: 2 subjects). The same outlier treatment was applied to the compensation measures, and outliers set to NA (Onset CC: 2 subjects, Onset V: 2 subjects, Coda V: 2 subjects, Coda CC: 3 subjects) as well as for the Perception data (PC1 Perception: 1 subject, PC2 Perception: 2 subjects).The data was divided into four datasets, each comprising one perturbation measure as the dependent variable (Onset CC, Onset V, Coda CC, Coda V) and the six principal components from the tapping tasks, as well as the two principal components from the perception tasks, and the motor variability of the unpaced tapping task (see Table 4.3 for an overview of measures). The four datasets were scanned for NAs, and each subject with more than three NAs in the data was removed from further calculations. Four subjects were removed from calculations based on this criterium (the same four as for the tapping blocks before the PCAs). The remaining NAs were replaced with *knn-Imputation* as performed on the raw data previous to the PCA. For nine subjects per dataframe, data was imputed for one to three values. After data exclusion and imputation, the remaining data comprised 28 subjects for onset CC compensation, 29 for onset V compensation, 27 for coda CC compensation, and 28 for coda V compensation.

Table 4.3: Overview of the measures of each of the three testing blocks along with the interpretation of the single PCs from the principal component analyses. The four compensation measures (shaded in grey) will, due to their difference in articulation, position within the syllable, and perturbation direction, always be treated as different dependent variables. Measures 1 to 3 from Tapping and Perception will serve as predictors in model fitting.

Test Block	Quality	Measure 1	Measure 2	(Measure 3)
Perception	Auditory Acuity	PC1: Auditory Acuity	PC2: Duration Discrimination vs. Beat-alignment Tasks	
Tapping	Motor Variability	PC1: Motor Variability	PC2: Continuous vs. Non-continuous Sound flow	spontaneous Tap
	Synchronization Accuracy	PC1: Synchronization Accuracy	PC2: Musicality	
	Synchronization Consistency	PC1: Synchronization Consistency	PC2: Musicality	
Perturbation	Onset Perturbation	Onset V compensation	Onset CC compensation	
	Coda Perturbation	Coda V compensation	Coda CC compensation	

4.3.3 Bootstrapping Model Predictors

Linear mixed models were fitted per compensation measure as the dependent variable (e.g., Onset CC compensation) and all of the tapping and perception measures as predictors, building the following model structure:

lm(compensation ~ PC1 Motor Variability + PC2 Motor Variability +
PC1 Accuracy + PC2 Accuracy +
PC1 Consistency + PC2 Consistency +
PC1 Perception + PC2 Perception +
Spontaneous Tap motor variability)

The number of predictors is relatively high for the number of observations in our dataset. To reduce the predictors to the most relevant ones, 100 random datasets per compensation measure were created with the same number of observations as in the original data. Accordingly, some subjects were excluded in a dataset, while others were included more than once. From those 100 datasets per compensation measure, 100 models with the above formula were fitted. This method called "bootstrap sampling" was invented in the seventies and is mainly used in machine learning to stabilize algorithms (Efron, 1992). In our case, bootstrapping many models with different subsets of the data aimed to avoid overfitting of the model and allowed for a proper justification for excluding or including specific predictors in the final model. For each of the 100 calculated models, R's *step* function selected the most relevant predictors for optimal model fit based on the Akaike information criterion (AIC). The predictors kept in more than 70% of the optimal models were included in modeling the original dataset. Table 4.4 summarizes how often a single predictor was kept in bootstrap modeling. Occurrences above 70% are shaded in grey and were used to model the original dataset.

Table 4.4: Percentage of inclusion for each predictor (columns) into 100 bootstrap models fitted after step selection per dataset (rows). Predictors included in more than 70% of the models were kept for modeling the original dataset and are shaded in grey. MV stands for Motor Variability.

Dataset	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2	spont. Tap
	MV	MV	Accuracy	Accuracy	Consistency	Consistency	Perception	Perception	MV
Onset CC	46	38	52	44	49	40	50	81	40
Onset V	77	60	55	41	51	66	89	34	74
Coda CC	85	57	49	66	49	47	89	93	99
Coda V	42	37	42	58	54	54	63	56	87

4.4 Results and Interpretation

The following section fits linear models with the predictors provided from bootstrap modeling to the original data. For each model, backward modeling with the *step* function provided the final model structure for which statistical outcome will be reported. The presentation of the statistical outcome will be followed by an interpretation and a brief discussion to allow for a smooth flow of understanding and classifying the results before they will be discussed more thoroughly in the discussion (section 4.5).

4.4.1 Main Results

4.4.1.1 Onset CC compensation

The original dataset with onset CC compensation values as the dependent variable was modeled with PC2 Perception as the predictor. Backward modeling with the step function confirmed PC2 Perception as a relevant predictor with a significant effect (estimate = 3.4, SE = 1.29, *t-ratio* 2.639, *p* = 0.0139*). Overall model fit was significant (adj. r-squared = 0.18, F-statistic = 6.966, df = 26, *p* = 0.0139*). The left panel of Figure 4.4 shows the onset CC compensation (y-axis) and PC2 Perception (x-axis). The reported relationship indicates that speakers who are better at perceptual beat-alignment judgments in music or speech compensate more to the introduced onset CC perturbation. This relation is likely tied to p-center perception: The p-center is a temporal landmark that serves for the isochronous organization of fluent speech (Morton *et al.*, 1976). This p-center lies in the onset of syllables, more precisely in the transitions between onset and vowel (Cummins and Port, 1998). Therefore, it is assumable that subjects who can precisely detect a mismatch of the synchronization point in beat-alignment tasks can also more easily identify and classify a p-center shift in the auditory feedback. The applied onset perturbation caused a shift of the p-center location by changing the relation between onset CC and following vowel, and speakers who could identify this shift more precisely compensated more for it.

4.4.1.2 Onset V compensation

The dataset with onset V compensation was modeled with PC1 Motor Variability, PC1 Perception, and spontaneous Tap Motor Variability as predictors. Backward modeling dropped the unpaced tapping task but kept the other two predictors resulting in an overall significant model fit (Adj. r-squared = 0.17, F-statistic = 3.9, df = 26, p = 0.03*). No significant effect was found for PC1 Motor Variability, but a significant effect for PC1 Perception (estimate = 3.4, SE = 1.3, *t-ratio* = 2.6, *p* = 0.015*). The middle panel of Figure 4.4 shows the relationship between compensation to the vowel in the Onset condition (y-axis) and PC1 Perception (x-axis). Thereby, more compensation to the vowel in the Onset condition is related to a better general auditory acuity performance. Analogously to the onset CC compensation relation with PC2 Perception, a better perception of the auditory mismatch might lead to more remarkable articulatory adjustments when the auditory feedback is altered. However, the beat-alignment abilities might not be most important anymore, as the vowel duration is determined at the end of the vowel where p-center location is already determined. At this point, higher general duration discrimination abilities might enhance the detection of a total duration mismatch of the vowel.

4.4.1.3 Coda V compensation

For understanding the compensation in the vowel of the Coda condition, only motor variability of the spontaneous tapping task was implemented as a predictor (and kept in backward modeling) with a significant effect (estimate = 3.36, SE = 1.428, *t-ratio* = 2.353, *p* = 0.0265*). Overall model fit was therefore also significant (adj. r-squared = 0.144, F-statistic = 5.538, df = 26, *p* = 0.0265*). The right panel of Figure 4.4 shows the relationship between compensation to the vowel in the Coda condition (y-axis) and motor variability of the unpaced tapping tasks (x-axis). The Figure presents compensation in %, meaning a positive estimate represents compensation, while a negative estimate indicates that the speaker followed the perturbation. For the coda vowel, however, recall that a compensatory response is actually realized with vowel shortening in production. More compensation to the coda vowel is here related to greater motor variability. The more variably subjects were tapping in the spontaneous unpaced tapping task, the more they compensated for the vowel in the Coda condition. This effect calls the general interpretation of differences in onset vs. coda compensation in chapter 2 to mind: While subjects compensated significantly for the onset vowel, coda vowel, and coda CC, no significant compensatory response was observed for the onset CC segment. Our main explanation for the lack of onset CC compensation was the articulatory stability of complex onsets as elaborated in the Articulatory Phonology/Task-Dynamics framework. Thereby, the structural stability of the onset also implies greater articulatory entrenchment and less malleability in the face of an auditory feedback perturbation. Accordingly, the coda segments provide more structural malleability by definition, and the data in the current section indicates that greater *individual* motor variability further enhances the malleability during auditory feedback perturbation.

One could ask why this relationship is found for the vowel in the Coda condition but not for the vowel in the Onset condition. As an answer to that, we want to stress once more that the focal temporal perturbation applied in this study always consists of two parts, the stretching and the compressing. The whole targeted sequence and the starting point of perturbation influence the nature of the reaction, as shown and discussed in chapter 3.

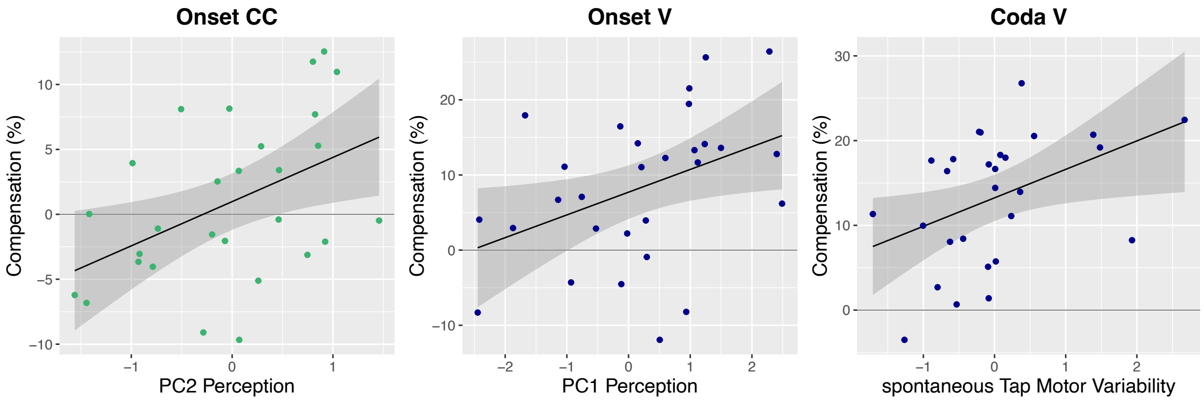


Figure 4.4: Relation between the significant predictors and the respective compensation measure (from left to right: Onset CC and PC2 Perception, Onset V and PC1 Perception, Coda V and spontaneous tapping).

4.4.1.4 Coda CC compensation

The four predictors PC1 Perception, PC2 Perception, PC1 Motor Variability, and spontaneous Tap Motor Variability were included in the linear model to understand compensation in the coda CC segment. All four predictors were kept in backwards modeling and had a significant impact (PC1 Motor Variability: estimate = 2.3, SE = 0.75, *t-ratio* = 3.073, *p* = 0.005**, PC1 Perception: estimate = -2.58, SE = 0.800, *t-ratio* = -3.219, *p* = 0.004**, PC2 Perception: estimate = 4.345, SE = 1.33, *t-ratio* = 3.263, *p* = 0.004**, spontaneous Tap Motor Variability: estimate = 5.14, SE = 1.2, *t-ratio* = 4.109, *p* < 0.001***). Overall model fit was quite a bit higher than in the previous models (adj. r-squared = 0.59, F-statistic = 10.23, df = 22, *p* < 0.001***).

Figure 4.5 shows the relation between the single predictors on the x-axis and compensation to coda CC on the y-axis. Firstly, the magnitude of compensation correlates highly with PC1 Motor Variability and motor variability of the unpaced tapping task. The directionality indicates that subjects with higher motor variability compensate more, as found before for the coda vowel compensation. This relationship is even stronger than for the coda vowel compensation, assuming the syllable coda segment to be even more affected by individual motor stability. Secondly, both Perception PCs correlate with the magnitude of compensation but with different directionality.

PC2 Perception (associated with better beat-alignment abilities) correlates positively with compensation, indicating that better beat-alignment judgments are related to more compensation. However, the factor loadings provided by Table 4.2 further indicate that better beat-alignment but *worse* coda duration discrimination abilities are captured in PC2. PC1 Perception correlates negatively with the coda CC compensation magnitude, suggesting that *worse* general auditory discrimination abilities are coupled with *more* compensation. For PC1, this relationship is the other way round than the previously reported relationship between PC1 Perception and compensation to the onset vowel, which will be further discussed in section 4.5.

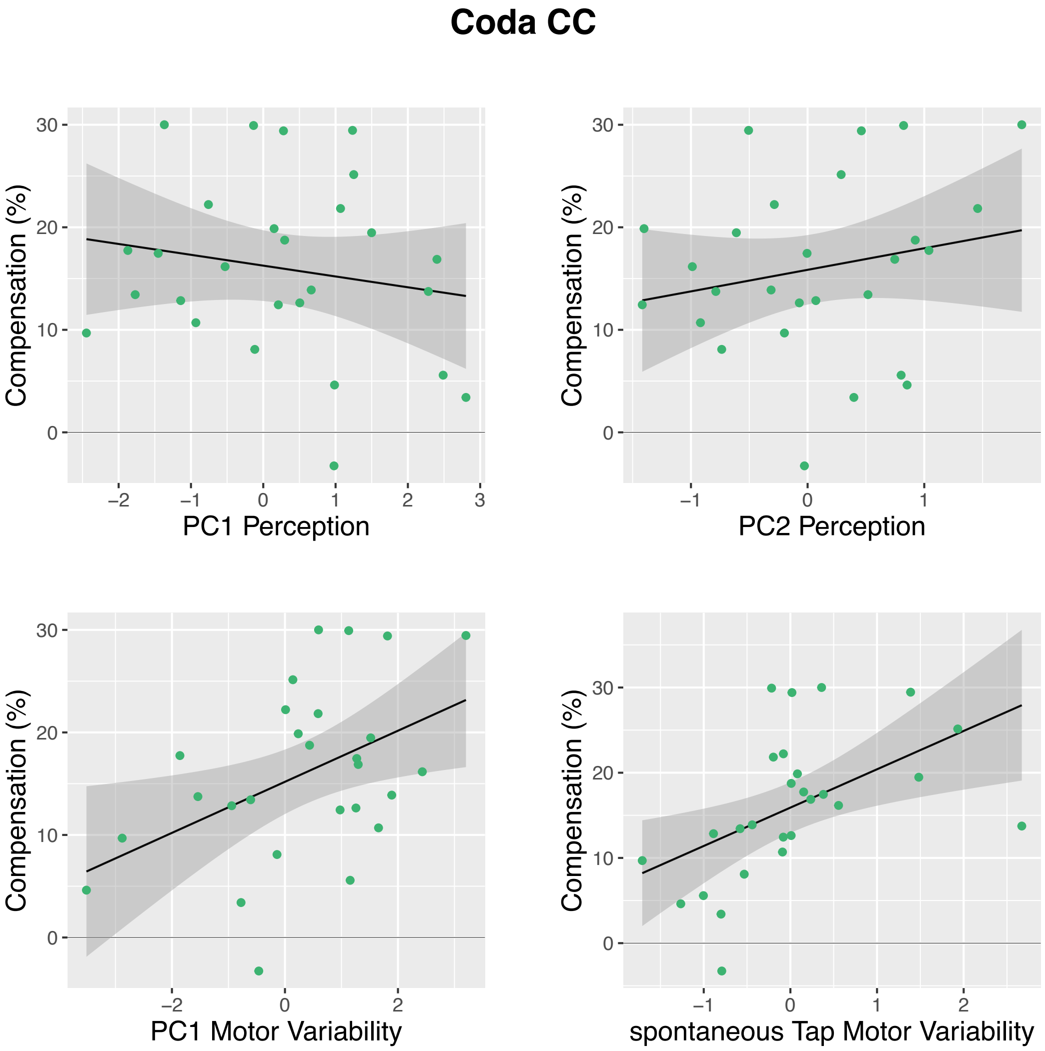


Figure 4.5: Relation of each of the significant predictors to coda CC compensation. Compensation on the y-axis, predictors on the x-axis. Upper left panel: PC1 Perception, upper right panel PC2 Perception, lower left panel: PC1 Motor Variability, lower right panel: spontaneous tapping).

The analyses in this chapter primarily aim at understanding and interpreting relationships in the data. However, the model fit for coda CC compensation with an adjusted r-squared of 0.59, indicates that quite a good behavioral prediction can be possible. Therefore, we suggest that the four predictors incorporated in our model could help understand the relationships between our three domains perception, tapping, and compensation and even predict coda CC compensation. Figure 4.6 depicts the actual measured compensation values (x-axis) against the predicted values from the linear model (y-axis). The visual presentation neatly substantiates the adequacy of the four predictors for predicting compensation to the temporally perturbed coda segment.

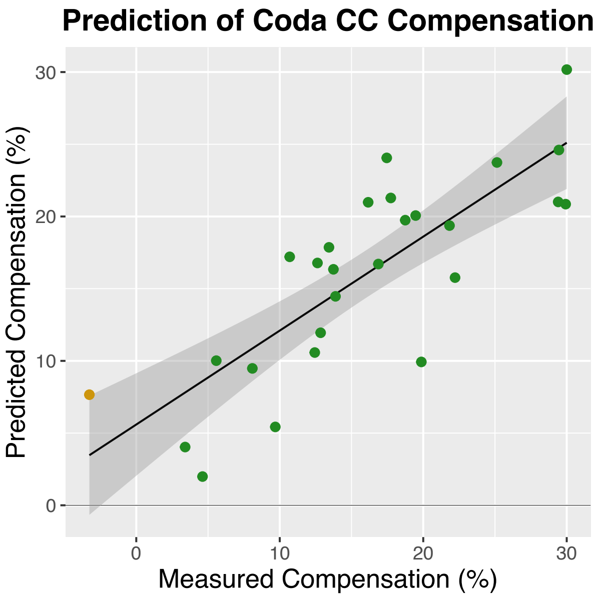


Figure 4.6: Predicted compensation (y-axis) and actual measured compensation (x-axis) to the coda CC perturbation as given by the linear model. Dots represent single subjects. One subject who followed the perturbation (measured compensation) is marked in gold and will be discussed further below.

4.4.2 *Further Investigations*

While the previous section gave a good idea of how perceptual and motor executive abilities relate to responses to temporal auditory feedback perturbation, two follow-up questions might have occurred to the reader: If tapping motor variability and perception both relate to coda CC compensation, do they correlate with each other? Furthermore, if *general* motor variability in finger tapping correlates strongly with malleability in the face of temporal auditory feedback perturbation, would this also be seen for *speech* motor variability? The following section briefly examines these two questions by correlating PC1 Motor Variability with PC1 and PC2 Perception. Subsequently, two different measures of speech motor variability will be presented and their relation to the four compensation measures examined. The additional analyses can contribute to a better understanding of the interaction between perception and motor action and give insight into how effects of general motor abilities also apply to speech motor abilities.

4.4.2.1 *General motor variability and perception*

Linear models were calculated between PC1 Motor Variability and PC1 Perception as well as PC2 Perception. All subjects with data in at least one of the four compensation data frames were included in the calculation (35 subjects). No relation between perception and motor variability was observed (see Figure 4.7).

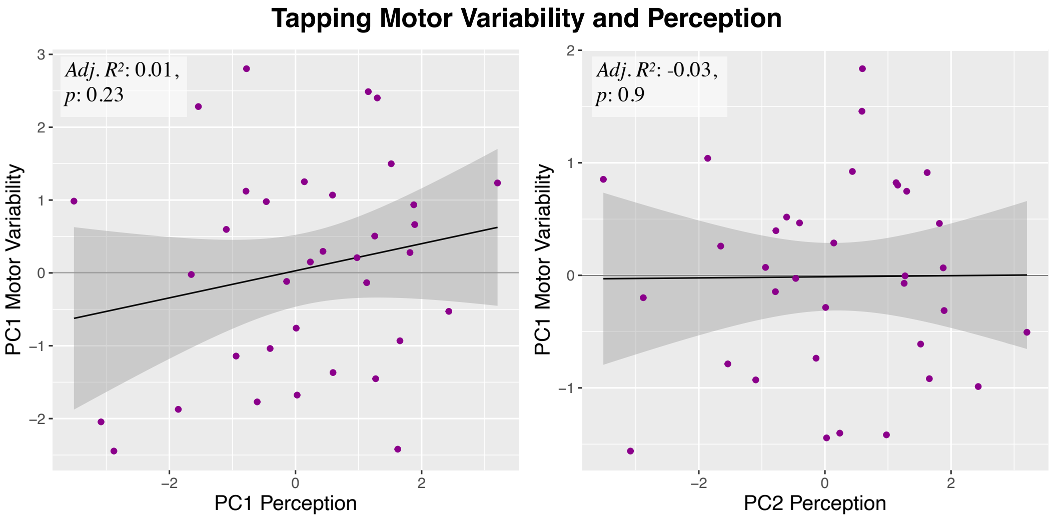


Figure 4.7: Relation between PC1 Motor Variability and PC1 Perception (left panel) and PC2 Perception (right panel). Effects are non-significant.
(right panel). Effects are non-significant.

4.4.2.2 *Speech motor variability and compensation*

For assessment of speech motor variability, we extracted two measures from two different production patterns. Firstly, we calculated the coefficient of variation (standard deviation divided by the mean) of the normalized segment durations (V and CC) produced in the baseline phase per experiment condition (Onset/Coda) per subject. The coefficient of variation was then correlated with each of the four compensation measures.

Secondly, we calculated the inter-vowel-onset interval (IVOI) between single words in a read wordlist to measure speech motor variability more similarly to the unpaced tapping task. This wordlist contained the stimuli from the wordlist in the tapping task (section 4.2.2.4) and was read by every subject before the perturbation experiment started to get the speakers used to speaking with the special in-ear headphones. The coefficient of variation (CV) of the IVOI was then also correlated with the four compensation measures. Unlike tapping motor variability, neither the CV of the baseline segment durations nor the CV of the IOI of the read wordlist were significantly related to the amount of compensation, as calculated by linear models and summarized in Figure 4.8.

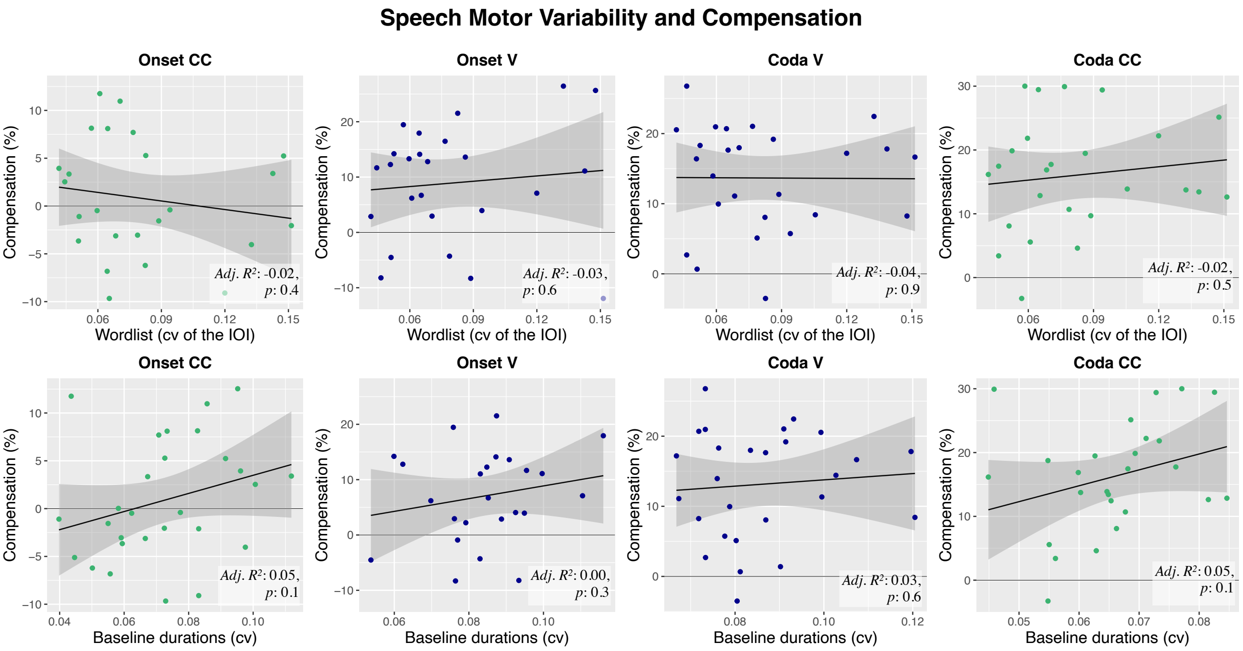


Figure 4.8: Relation between speech motor variability and compensation. Compensation on the y-axis and speech variability measures on the x-axis. Speech motor variability measures examined from wordlist reading are shown in the upper panels, production variability of baseline durations from the auditory feedback experiment are in the lower panels. Effects are non-significant.

4.4.3 *Summary of Results*

The previous sections outlined the relationships between compensation to the two segments (V and CC) in the two different perturbation conditions (Onset/Coda) to tapping performance and perceptual acuity. The analyses showed that perception and tapping motor variability are most relevant for predicting compensatory reactions. Synchronization accuracy and synchronization consistency did not seem beneficial for understanding the relationship between the tapping performance and compensation and were excluded from further calculations or considerations. For reactions to the onset perturbation, perceptual abilities seemed to be the most promising predictor, while in the coda perturbation, individual motor variability was highly related to the responses. By far the most robust model performance (based on the adjusted r-squared) was retrieved from modeling the reactions to the coda CC perturbation, with both motor variability and perception contributing significantly to predicting the response. Additional measures showed, however, that tapping motor variability was not related to perceptual abilities. Further, the examination of motor variability retrieved from speech stimuli showed no relation to compensatory behavior. The two perception components contributed in different ways to the models and therefore remain complex and need further discussion.

4.5 Discussion

The presented investigation explored different temporal qualities of motor action and perception with and without synchronization components. The extracted qualities were summarized in terms of their underlying dimensions (principal components) and further served as parameters for finding relationships with the responses to temporal auditory feedback perturbation. The following discussion will more thoroughly debate the results from this chapter. Since this is an exploratory study, the results have to be interpreted cautiously. Interpretations that aim at a general understanding of speech timing mechanisms and conclusions that can be drawn by including the findings of the previous chapters will be postponed to the general discussion in chapter 5. The general discussion in chapter 5 will also include an outlook for further studies that can build on the findings from this and the previous chapters.

4.5.1 *Compensation, Perception, and Motor Variability*

With different perception tests, our data captured two main qualities of perceptual performance: on the one hand, general auditory acuity (PC1 Perception) and, on the other hand, beat-alignment performance (PC2 Perception). For CC and the vowel in the Onset condition, better perceptual abilities were associated with more compensation. Thereby, perceptual beat-alignment performance was the relevant predictor for onset CC compensation (which was interpreted as tied to p-center perception), while for the vowel in the Onset condition, general auditory acuity proved relevant. This relationship indicates that the better speakers discriminate temporal discrepancies, the more they compensate for a temporal mismatch introduced by auditory feedback perturbation. In chapter 2, we concluded that onsets form the anchor for temporal alignment of fluent syllables following mechanisms of event-based timing, while other parts of the syllable instead follow duration-based timing mechanisms.

The coupling of more CC onset compensation with a better perceptual beat-alignment performance and more vowel onset compensation with better general perceptual acuity strengthens this assumption. However, as compensation to the CC onset segment was not consistently found but instead both compensatory and following responses, it is not entirely clear if a strong negative response (i.e., following) can simply be counted as "very little" compensation. If we only look at compensatory responses (the positive estimates) for both discussed relationships (Figure 4.4 left and middle panel), the linear relationship between the perceptual and the compensatory component would in fact probably look very similar.

This finding extends the knowledge of previous research, where more compensation to spectral alterations was found in speakers with higher spectral auditory acuity (Villacorta *et al.*, 2007; Brunner *et al.*, 2011). The contribution of different qualities in speech perception to the responses indicates that temporal perceptual ability is complex. While, e.g., in Villacorta *et al.* (2007), along with spectral manipulations of F1, the auditory discrimination ability of F1 was assessed, the underlying dimensions of temporal planning and execution are not self-evident (see chapter 3 for discussion). Our examination showed that two perceptual dimensions are connected with the temporal organization of speech, namely duration discrimination abilities and beat-alignment judgments. While the perception components seemed relevant for the responses, none of the extracted tapping qualities showed a relation to the compensatory responses in the Onset condition.

In the Coda perturbation condition, tapping motor variability correlated with the vowel and the CC segment, whereby greater motor variability was connected to more compensation. This finding contributes to our understanding of segment stability and malleability when the auditory feedback is altered: Syllable onsets naturally show more entrenched *structural* articulatory stability resulting in less malleability in auditory feedback alterations. The malleability of nucleus and coda segments can further be shaped by *individual* motor/feedforward stability. Less motor variability indicates a more stable feedforward system which is less fragile when one of the feedback channels provides information requiring a deviation from the stable feedforward pattern.

Both perturbed segments that correlate with motor variability share the occurrence of adaptation effects (that is, remaining compensatory adjustments when regular feedback is restored, see chapter 2.4.3). This link couples general motor variability tightly to timing mechanisms in speech planning and feedforward representation rather than to online control mechanisms. The predicted coda CC compensation shown in this chapter in Figure 4.6 further suggests that motor variability affects the deviation from the original motor plan, no matter in which direction (following or compensatory response): The one subject that produced negative measured compensation (followed the perturbation) is marked in Figure 4.6 in a golden color. However, the predicted compensation for this speaker is similar to a positive response of the same amount, meaning that the predicted value corresponds to the predicted value of an equally strong compensatory (positive) response. Since this applies only to one speaker, this prediction should not lead to conclusions. Still, it should give an incentive to think differently about negative responses in motor activity terms than in perceptual terms: While following the perturbation increases the auditory mismatch, the motor deviation from the feedforward plan is equally strong, no matter if a speaker compensates by, e.g., 10% or follows the perturbation by 10%.

In the coda CC segment, aside from motor variability, both PCs of Perception contributed significantly to overall model fit. Thereby, better general discrimination abilities (PC1 Perception) lead to *less* compensation, while a better performance in perceptual beat-alignment tasks leads to *more* compensation. This composition somewhat diverges from our expectations since the beat-alignment performance was previously characterized as a quality most essential for aligning *onsets* of syllables in fluent speech (and therefore affecting onset CC compensation).

One explanation for this directionality of effects can be seen when concentrating on the coda perception task, which shows high negative factor loadings for PC1, but even higher positive factor loadings on PC2 (see table 4.2). Considering this, the contribution of both Perception PCs indicates that a *poorer* temporal discrimination ability of manipulated coda segments is associated with *more* compensation to the manipulated coda CC segment.

This effect could be attributed to the experience of self-agency during auditory feedback perturbation (Korzyukov *et al.*, 2017): Small feedback shifts that do not deviate massively from predictions lead to compensation, while large feedback shifts suggest that the feedback was not self-generated and might lead to a greater reliance on internal predictions (Subramaniam *et al.*, 2018). For speakers with higher auditory acuity, a feedback shift might lie more likely outside the area of possible self-generated errors, leading to less compensatory response and a greater reliance on internal models. The stability of the feedforward prediction, however, can also vary between individuals. Thereby, speakers with less stable internal models compensate more. Altogether, speakers with relatively unstable feedforward predictions compensate more, and speakers with greater perceptual auditory acuity rely more on these internal representations when the auditory shift is large enough.

Further, when regular auditory feedback is restored in the aftereffect phase, low auditory acuity speakers might still compensate more because they do not necessarily classify the altered productions as a mismatch from the auditory representation. The additional examination of a connection between general motor variability and perception shows that motor stability and perceptual abilities do not necessarily indicate a relationship (section 4.4.2.1 and Figure 4.7). Accordingly, all possible combinations of auditory acuity performance and motor variability levels can be expected.

This finding is in line with a very recent study by Cheng *et al.* (2021). They looked more deeply into speech variability and perceptual abilities by measuring variability of vowel repetitions in production, the so-called centering ratio from the start to the end of a vowel in production (Niziolek *et al.*, 2013), distinctive vowel contrast in production, and categorical labeling consistency of a distinctive vowel contrast in perception (Cheng *et al.*, 2021). They did not find any relation between either of the three production variability measures with the perceptual labeling component, indicating that the perception–

production relation is not straightforwardly determinable. This finding further supports our assumption that many different structural and individual aspects contribute to the perception-production interplay.

It is still unclear why more compensation to the Onset condition is associated with better perceptual abilities, but for compensation to the Coda condition, this relationship is inverted (at least when focusing on the coda perception component). Recapitulating, we have seen that knowledge gained from spectral auditory feedback alterations cannot be transferred one-to-one to responses to temporal auditory feedback manipulations. Moreover, we have seen that the prosodic properties of fluent speech shape the responses. Considering both, it is genuinely assumable that auditory feedback plays a different role in reactions to temporal alteration of onset and nucleus than to nucleus and coda. With the exploratory paradigm of this study, further interpretation/conclusion about the results would not be wise and will therefore be postponed pending further investigations.

4.5.2 Speech Motor Variability and Compensation

While *general* motor variability in tapping correlated with compensatory responses in the coda perturbation condition, a similar relationship could not be found for measures of *speech* motor variability. By closely examining previous studies, this result turns out to be in line with investigations on spectral variability and compensation to spectral perturbation.

In spectral studies, the amount of compensation correlated with the variability of distinctive contrast production (Ghosh *et al.*, 2010; Brunner *et al.*, 2011; Franken *et al.*, 2017), but not with the variability of one single parameter (e.g., F1) in repeated phoneme productions (MacDonald *et al.*, 2010; MacDonald *et al.*, 2011). Mixed findings were also provided by Nault and Munhall (2020) who conducted a study on inter- and intraspeaker variability. They measured the standard deviation of the first two formants of vowels produced in the baseline phase of a spectral perturbation experiment and found a relation between F1 variability in the baseline and F1 compensation in the hold phase, but no contribution of baseline variability of F2 as a predictor for compensation to perturbed F2.

In a model incorporating baseline variability, perceptual acuity, and vowel space as predictors for compensation, neither F1 nor F2 baseline variability was a significant predictor, but perceptual acuity was. Nault and Munhall (2020) attributed the different results in the literature to undersampling, whereby single subjects that show an outstanding reaction contribute enormously to the model outcomes. However, another recent study neither found relations between adaptation and vowel spacing in the baseline phase nor correlations between adaptation and variability in productions of single baseline phonemes (Parrell and Niziolek, 2021). In this view, the non-existent relationship between speech variability and compensation in our data is in line with the findings by Parrell and Niziolek (2021), MacDonald *et al.* (2010), MacDonald *et al.* (2011), and partially Nault and Munhall (2020).

In our data, the examination of variability in reading the wordlist could be improved by calculating the exact p-center positions rather than calculating inter-vowel-onset-intervals, and the reading instructions could be more precise. In our case, all the subjects read the wordlist in an isochronous style. The instruction, however, did not explicitly incorporate the order to speak as regularly as possible. Further, other measures for speech motor variability should be considered for future studies than the two approaches presented in this study. Analogously to studies that measured spectral distance variability between sounds, the vowels /a/ and /a:/ in German differ almost exclusively in duration and would allow for measuring the variability in duration-based phoneme *contrast*. This contrast-based variability could be correlated with compensation magnitude to temporal perturbation to further contribute to the preceding discussion.

However, it has once again to be kept in mind that temporal information of speech is different from spectral information: While spectral properties of fricatives and vowels serve to distinguish similar sounds from each other, duration's primary purpose is not to distinguish sounds but to give their spectral evolvement a stage. The idea of individual variability in the distinctiveness of targets might not be relevant when the temporal change of one of the perturbed sounds does not necessarily result in another phoneme, or, in the perturbation of /a/ in chapter 2, does not result in another word. The distinctive function of duration is much less pronounced than the distinctive function of spectral properties of speech. Duration and timing are certainly not arbitrary but follow different goals, such as enabling fluency and intelligibility and realizing prosodic aspects of speech.

4.5.3 *Outlook*

The current study incorporated many tests to get an overview of connections in human speech and non-speech performance. The three-way connection between perception, motor action, and compensation to temporal perturbation offers much scope for future investigations. In particular, motor variability of tapping tasks seems worth exploring further; for this, unpaced and paced tapping should be considered. The performance in tapping to music did not necessarily improve our understanding of speech and non-speech motor activity. The discrimination performance for temporally manipulated perception stimuli contributed to predicting compensation to temporal auditory feedback perturbation. Thereby, the temporal manipulation of speech stimuli proved to be interesting, and so did speech and music beat-alignment tasks. The duration discrimination of pure tones was not overall beneficial and could be omitted in further studies to reduce the number of tests and improve the perception test paradigm instead. For example, this study made use of a 2-interval 2-alternatives forced-choice staircase paradigm, which entails deficits such as the need for catch-trials. However, this paradigm was chosen because it allowed for conducting many tests and was short enough to test children. The children's data were not presented in this study but were assessed with the same test battery for other research reasons, which limited the possibilities of the design complexity. In follow-up investigations, a 4-interval 2-alternative AABA design would provide more reliable threshold estimations than the 2-interval paradigm or an ABX paradigm (Gerrits and Schouten, 2004). The current study provided many advantages for future studies to make a more targeted selection of perception and tapping tests. Further, the findings gave a firm understanding of how feedback and feedforward mechanisms in speech and non-speech are connected and shape timing strategies in speech production.

Chapter 5
General Discussion and Outlook

This thesis presented three investigations that contribute to our understanding of speech timing mechanisms with interacting feedback and feedforward systems. Chapters 2 and 3 examined prosodic timing patterns under temporally perturbed auditory feedback, with the finding that speakers indeed compensate but with different response patterns that depend on the underlying prosodic structure. The fourth chapter tested the contribution of individual auditory acuity and general motor stability to the perturbation response data from chapter 2. The data indicated that both perceptual acuity and motor stability affect responses but with an additional influence of the perturbed segments' structural stability.

5.1 Summary of the Main Findings

5.1.1 Compensation and Adaptation

The first main finding of this thesis is that speakers compensate for locally applied temporal real-time auditory feedback perturbations in both directions, meaning with lengthening and shortening of segments in production. This bidirectional reaction pattern could be observed in the second chapter when speakers lengthened a compressed vowel (Onset condition) and shortened a stretched vowel (Coda condition). Bidirectional compensation to real-time perturbation of speech timing has not been observed before (only one direction was found in Floegel *et al.*, 2020) and indicated that the temporal organization is monitored via auditory feedback. Adaptation to the perturbed segments in the Coda condition further indicated that representations can be updated eventually but as a function of the syllable position and the onset of perturbation.

5.1.2 *Onset Stability*

The second main finding was that syllable onsets are more entrenched in the motor system than syllable codas. The absolute temporal extent (in ms) of syllable onsets was resistant to temporal auditory feedback perturbation in both chapter 2 and chapter 3. In chapter 3, onset stability was observed independently of syllable position within the word given that speakers did not change the temporal production of complex onsets in either a word-initial unstressed or a word-medial stressed syllable. This effect was also attributed to the fact that in onsets, less information about the relative timing of segments within the greater prosodic unit (e.g., the syllable or the word) can be determined by the auditory feedback in the online control. The temporal stability further suggested that onsets set the grid for temporal alignment of following productions within the sequence. However, in chapter 3, adjustments in later parts of the word elicited patterns of compensatory response for the perturbed onset segments in both targeted syllables when viewed on the word-level, indicating that speakers aim for relative timing patterns controlled by the auditory feedback.

Similar in both perturbation experiments is the behavior of the single consonants in the perturbed onsets. While the whole CC sequence remains constant in absolute durations, the leftmost consonants (/p/ in chapter 2 and /t/ in both conditions in chapter 3) rather shorten (compensatorily) while the rightmost consonants (/f/ in chapter 2 and /ʃ/ in both conditions in chapter 3) rather lengthen (following the perturbation). The right-most consonant thereby follows the adjustment direction of the vowel. This systematic response pattern calls the principles of coupled oscillators to mind, whereby the two consonants in the onset are anti-phase coupled with each other (Goldstein *et al.*, 2009). Further, it remains questionable what in this case happens to the overlap in the onset, and if and how the c-centers and p-centers are shifted. In complex CC onsets, the first consonant undergoes a left-ward shift and the second consonant a right-ward shift relative to the c-center. The perturbation seems to increase this coordinative pattern in production. The sequential acoustic measurements of single segments can certainly not give insight into these questions, and might even disguise effects of gestural reorganization. The true coordinative mechanisms behind this can only be revealed with kinematic data.

5.1.3 Lexical Stress and Syllable Position

The second experiment further aimed at extending our knowledge about prosodically shaped timing mechanisms by examining temporal alterations of lexical stress. Responses to the perturbation of a stressed and an unstressed syllable were compared, whereby in the Stressed condition, the perturbation weakened the stress pattern. As expected, speakers compensated more in the stressed syllable to maintain the desired stress pattern. Most remarkably, however, the global timing of all segments within the word differed between perturbation conditions. The perturbation of the word-initial syllable caused a global slowing down of the following segments, while the perturbation of the stressed syllable caused local temporal adjustments within the targeted syllable. Therefore, the third main finding is summarized with the assumption that both syllable position and stress pattern shape the responses to temporally altered feedback. Thereby, the nucleus of an unstressed syllable does not allow for a tremendous temporal extension (especially not in direct comparison with a stressed vowel of the same category), and therefore compensation must be spread bit by bit over following segments. The target of a stressed vowel, on the other side, has less strict limitations towards longer durations and allows for a significant lengthening of the vowel without the need for changing other segments drastically. In the discussion of chapter 3, the conflation of stress and accent in the Stressed condition was classified as a limitation of the study. However, it remains questionable how these two concepts can be separated in natural speech. At least in German, the accented syllable will always be a stressed syllable, and the stressed syllable of a target word most likely carries the accent in a neutral context. In this view, other languages would be worth investigating, such as Polish where word stress is always on the penultimate. Another prosodical interesting and more complex example to investigate would be Finnish. Finnish has segmental quantity oppositions in both consonants and vowels which limits the possibilities of duration to signal prosodic distinctions. In Finnish, the first syllable always carries word stress. Accent, on the other hand, depends on the moraic structure and is realized with a rising f0 on the first mora and a falling f0 on the second mora (Suomi *et al.*, 2003). Further than that, there is a complex relationship between syllable and segment duration in stressed syllables depending on moraic pattern, syllable structure, and voicing (Suomi and Ylitalo, 2004).

Saying this, the combination of accent and stress in chapter 3 should rather be seen as a point of departure for research on temporal alterations of prosodic cues, opening up a wide range of possibilities for future investigations.

5.1.4 Different Findings from the two Perturbation Experiments

The experiments in chapters 2 and 3 outlined prosodically shaped timing mechanisms and their stability in the face of a temporal auditory feedback perturbation. The results of the two perturbation experiments differ in the finding that CC durations in chapter 2 do not show any temporal adjustment in absolute or word-normalized durations. In contrast, in chapter 3, word-normalized durations of the onset CC segment indicate a compensatory response of timing (but absolute durations do not). The compensatory response is thereby achieved by adjusting all following segments in the word. Close examination of the differences between the two experiments leads to the suspicion that the speech material has likely conditioned the varying outcome in word-proportional response: While chapter 3 investigated a three-syllabic word ("Tschetschenen"), chapter 2 examined a three-syllabic compound, whereby the first syllable was the perturbed one and built one stem of the compound ("**Pfann**kuchen", first word stem in bold). The temporal organization might, in this case, instead aim at realizing the duration patterns of the single stem rather than of the whole compound. Thereby, the non-compound "Tschetschenen" provides more segments to reorganize the timing pattern when the initial onset is perturbed compared to the word stem "Pfann" in "Pfannkuchen". In the perturbation of the first syllable in "Tschetschenen", all following segments were altered to "catch up" to the perturbation, and in the second stressed syllable, the stressed vowel /e:/ served to catch up to the introduced perturbation. In "Pfannkuchen", the more weakly stressed vowel /u:/ and the coda /n/ in "kuchen" did probably not adjust (enough) to change the relative timing of the onset CC.

5.1.5 Motor Stability and Auditory Acuity

The fourth main finding, derived from chapter 4, is that both individual perceptual and individual motor abilities contribute to the speaker's response to temporal auditory feedback perturbation and therefore both influence speech production. While the contribution of perceptual abilities was complex and differed between syllable segments, the examination of motor variability brought the novel finding that in speech timing, not only *structural* stability of syllable segments but also *individual* motor stability affects the malleability of productions when temporally altered feedback is provided. Speakers who showed less stable motor executions in non-speech tasks were found to compensate more to the perturbation, indicating that the whole motor system is more malleable to perturbations. The segments that experienced adaptation were more tightly associated with individual motor stability (Coda condition segments), while segments that experienced online compensation/reactive feedback control were more tightly associated with auditory acuity (Onset condition). This connection underlines one conclusion already made previously in this thesis: there must be different weightings of feedback and feedforward systems in speech production. However, no relationship was found between speech variability and responses, which might be due to the assessment of speech variability. For a more thorough insight into this connection, the perturbation data from chapter 3 will be also linked to individual motor execution and auditory abilities in future studies. In chapter 4, the use of many tasks and predictors increased the risk of data loss due to incomplete/missing data in single tasks. Since many statistical approaches cannot deal with incomplete datasets, many participants with missing data in only one predictor have to be removed, or other solutions need to be found, such as knn-imputation as used in this study. However, we believe that the study provides a valuable foundation for guiding future studies in the selection of more targeted perception and tapping tests for similar research approaches. In particular, motor variability in tapping tasks seems worth further exploration; especially unpaced tapping as a measure of general internal motor stability should be considered. Further, a smaller selection of tasks would allow for more repetitions per single task and hence multiple datapoints per participant, which opens up more possibilities for other statistical approaches, such as linear mixed models.

5.2 Modeling Speech Production

5.2.1 *Target Dimensions*

The main findings reported in the previous section suggest that auditory feedback contributes to the establishment and control of temporal properties of speech. But how can a temporal representation be determined? Do speakers establish temporal speech targets similarly to spectral targets of speech? Solely the fact that speakers compensate and adapt for temporal feedback shifts could insinuate that there must be targets speakers aim for and that temporal properties are similar to spectral properties flexibly guided by the auditory feedback. Nevertheless, the dimension of the target might not be easy to determine. The analyses in chapter 3 revealed that deviations from natural productions on the phoneme level (in absolute durations) cause different proportions when viewed on the word level.

Previous research suggested that the duration of speech segments is rather controlled for a proportional relationship of the segments in a larger unit, such as a syllable, than for each segment (Fowler, 1981b; Munhall *et al.*, 1992; Munhall *et al.*, 1994; Mitsuya *et al.*, 2014). This conclusion was also drawn in this thesis by the findings in chapter 3 when local compensation patterns were achieved with global responses of all segments to match the appropriate relative duration within the word (section 3.3.3). This finding underlines an (auditory) target of relative durations within a higher prosodic unit. The lack of onset adjustment in both chapter 2 and chapter 3 in absolute durations, however, indicates a stable motor target of the onset segment that is pre-planned. If the auditory percept of this motor execution does not fit the prediction, however, the following segments can be altered, in some cases eliciting another speaking rate. Hence, it is assumable that there are temporal motor targets that are more or less stable depending on the prosodic structure. The overall timing of sounds in higher prosodic units is defined by a relative target that can be adjusted via the auditory feedback and is further influenced by the rate of speaking. Therefore, closely related to the temporal target dimensions is the question of how the speaking rate is planned.

An overall lengthening of segments in reaction to a stretched onset CC cluster (chapter 3, Unstressed Condition) indicated that the speech rate is adjustable via auditory feedback.

The temporal perception of the stretched CC segments thereby seemed to trigger a speaking rate that allows for compensation for the perturbed segments on the higher prosodic level. Not much research has targeted the question of how pre-planned speaking rate might be in the speech planning process (but see Rodd *et al.*, 2020). However, the planning process is supposed to happen as late as possible (Kello *et al.*, 2000; Damian and Dumay, 2007) and, indicated by our data, might even be adjustable after production onset in the online control of ongoing speech.

5.2.2 *Differences in Measuring spectral and temporal Compensation*

Multiple times throughout this thesis, we concluded that findings from spectral auditory feedback perturbations could not simply be transferred to temporal perturbations. Differences in the perturbation paradigm and the perturbed dimensions, that is, spectral shape or duration, require attention on all levels. The analyses in this thesis have opened up quite a few issues that underline why temporal parameters of speech need to be considered differently than spectral speech properties. One of these issues is how to measure compensation in the temporal domain, which is done differently in this thesis than in many spectral perturbations.

In spectral auditory feedback perturbation, compensation to a formant frequency is often measured as the opposite deviation to the applied shift (see, e.g., Niziolek and Guenther, 2013), whereby the size of the shift is previously defined (e.g., 200 Hz). However, this calculation did not seem directly appropriate for the temporal examination in our studies because the absolute amount of perturbation was applied as a function of the initially produced segment durations and therefore varied between speakers. Further, temporal perturbation is always two-sided by stretching and compressing the signal. To account for this, compensation measures that included the whole targeted sequence were computed, including the amount of perturbation (sections 2.4.2.2 and 3.3.2). Additionally, the production deviation as a percentage from the initially produced baseline productions was measured (see sections 2.4.2.1 and 3.3.3). The second mentioned measurement neatly illustrates the dimensional difference between the spectral and temporal scale: When a speaker in our study produced a vowel of 300 ms that was compressed in perturbation, and as a reaction to the perturbation, she/he lengthened the vowel production by 150 ms, we can easily conclude that she/he produced 150% of the initial vowel duration reporting the difference in *quantity*. Thereby, the amount of motor action triggered by the perturbation is captured.

On the other hand, when a speaker produces a vowel with an F1 frequency of 600 Hz, and in reaction to a spectral shift she/he lowers F1 to 300 Hz, one would not say she/he produced "half the formant (frequency)," but another *quality,* independent of the direction of adjustment[2]. The compensation measure then explains how the applied perceptual mismatch is accounted for within the acoustic vowel space and how acoustic goals are pursued. To summarize, the human capacity of perception differs between the Hertz and the time scale. Even though both scales allow for bidirectional change towards higher and lower frequencies or longer or shorter durations, both changes in frequency will be perceived as *different,* while the changes in duration can be perceived as either shorter or longer. Thereby, in spectral alterations, the adjustments relative to the applied shift might be most important, while in the temporal domain, the adjustments relative to the initial production give more information about the reaction. This leads to the assumption that spectral adjustments naturally aim primarily for an auditory spectral target, while temporal adjustments are more firmly governed by the initial motor plan for the desired sequence.

This assumption is further encouraged by the idea that spectral information itself is not directly represented in the motor system but rather the actions that lead to desired spectral targets, with the auditory feedback serving to arbitrate between the intended and actual spectral shape of a sound. In temporal coordination of speech, the temporal unfolding of gestures as defined in the motor plan generates speech timing.

5.2.3 *Current Speech Production Models*

To date, none of the existing speech production models can thoroughly explain the results of the investigations presented here. Our data show that both prosodically determined timing mechanisms and auditory feedback contribute to fluent speech production. In current modeling approaches, there are some models in which the auditory feedback

[2] Assume in this example that the listener is not a trained phonetician.

pathway is not explicitly modeled, and on the other hand models in which duration and coordinative timing are not modeled dynamically. For example, in Articulatory Phonology/Task-Dynamics, monitoring the coordination of gestures after their activation is not foreseen, and in DIVA, the temporal coordination of speech sounds with respect to prosodic structure is not explicitly determined.

The previous discussion indicated that the prosodic structure of the speech sequence and individual abilities in motor execution and perception shape the sensitivity to temporal auditory feedback shifts. The DIVA model accounts for a disparity in the use of auditory and somatosensory feedback. Thereby, auditory feedback is used earlier in the acquisition of speech segments, and later on, speakers establish individual preferences for the weighting between auditory and somatosensory feedback (Lametti *et al.*, 2012; Patri *et al.*, 2019). Further, Hickok (2012) assumes that auditory feedback is used to a greater extent to correct motor plans on the syllable level, while somatosensory feedback serves to control the sound-related speech gestures (Kröger *et al.*, 2020). This adds complexity to the previously raised question of the unit of temporal control and encourages our assumption that auditory feedback contributes more or less to the temporal control of segments in different structural positions in fluent speech. Assume that onsets rely more on somatosensory feedback than on auditory feedback. Then in auditory feedback perturbation, there is no error perceived between somatosensory prediction and somatosensory feedback. This assumption is supported in our data as there were no adjustments in motor response to perturbations of onsets. Further, segments in the onsets overlap to a greater degree than segments in the coda. This overlap per se is not encoded in the auditory signal but in gestural coordination, indicating a more significant role for somatosensory feedback on the one hand and feedforward readout on the other hand in sequences with more overlap.

Our data indicate that beyond a weighting of sensory feedback channels, the use of feedforward readout over feedback control depends on individual preferences and abilities, the phonological system, and the timescale of events. The latter refers to the fact that auditory feedback is naturally delayed, and in a word- or syllable-initial position, the auditory feedback cannot be used to the same extent for the online estimation of predicted timing as it is used for later parts in the sequence. Somatosensory feedback control of speech gestures, on the other hand, can be used for (close to) real-time corrections (Hickok, 2012).

This conclusion could be drawn from our data regarding temporal auditory feedback perturbations. However, the spectral perturbation study by Shiller *et al.* (2009) showed that speakers adapt for perturbed fricatives in the onset. This discrepancy, however, should not be seen as a contradiction to our results but rather once more as an indication that temporal perturbations need to be viewed impartially. Unlike consonant timing, there are no such structural differences suspected for the spectral shape of fricatives in onsets or codas. Further, to date, only the above study perturbed fricatives in the onset and another one fricatives in the coda (Klein *et al.*, 2019) in real-time. In the data of the latter study, the responses varied to a large amount between speakers and indicated different underlying strategies. Certainly, more research needs to be conducted that probes the responses to fricative perturbations or manipulations of other consonants by controlling the prosodic context. Especially sibilants require a precise motor strategy to form a specific constriction, thereby naturally generating more informative somatosensory feedback than vowels. Therefore, individual preferences in feedback channels and differences in motor stability could also shape the responses to the perturbation of sibilants.

5.3 Future Investigations

The two presented perturbation studies, and the combination of response data with non-speech motor and perception tests constitute highly novel investigations that contribute significantly to our understanding of timing mechanisms in speech. The findings of the three presented studies and their limitations should pave the way for subsequent investigations on articulatory malleability, feedback-feedforward interaction in fluent speech production, and the transferability of non-speech abilities on to speech production.

5.3.1 *Unexpected vs. Expected Perturbations*

One aspect that remains understudied in temporal perturbation is the precise effect of unexpected, random feedback shifts (compensation paradigm). In this thesis, both temporal perturbation experiments followed an adaptation paradigm, which led to reactions in the online control of speech timing (which encompasses both online compensation and reactive feedback control) and allowed the examination of adaptive behavior. Especially in chapter 3, however, it was not easy to disentangle which effects are adaptive or exclusively online corrections. Following the study by Cai *et al.* (2011), it remains unclear how speakers might react to an unexpected perturbation when the total duration of a segment is focally altered. One assumption about the limitations of the unexpected paradigm was already given in chapters 2 and 3: online compensatory shortening as a reaction to stretched durations in the auditory feedback should not be possible in real-time. Chapter 3, however, showed that compensatory shortening of a segment could be achieved in real-time (which in this case means within the same trial) by adjusting the following segments in the word, which was close to the findings by Cai *et al.* (2011). The introduction of both unexpected feedback alterations and consistently applied auditory shifts in the same speech material could give more profound insight into reactive control mechanisms and the ability to update timing patterns in speech. Conducting both paradigms with the same speech material could help more precisely disentangle the mechanisms that drive a particular response.

5.3.2 *Different Populations*

Another promising field for future investigations brings different populations to the fore. For studying different aspects of human behavior, examining populations that show an impairment or a systematic improvement in that particular aspect can contribute significantly to the understanding of underlying mechanisms and functions. Regarding speech timing, people who stutter show systematic deficits in producing fluent speech. Research has shown that people who stutter tend to compensate less for spectral alterations (Cai *et al.*, 2012; Cai *et al.*, 2014). This effect was mainly attributed to different incorporation of auditory feedback into the online control of ongoing speech in people who stutter (Cai *et al.*, 2012; Cai *et al.*, 2014), whereby an overreliance on auditory feedback in onsets might lead to repetitions (Civier *et al.*, 2010). Studying responses to temporally altered auditory feedback in people who stutter could give insight into the incorporation of auditory feedback into speech production and natural or impaired dominances of one mechanism over the other (e.g., auditory feedback, somatosensory feedback, or feedforward mechanisms). This investigation could benefit again from a focus on syllable structure, to investigate differences in response to a perturbation in syllable onsets as compared to syllable codas between people who stutter and fluent speakers. With respect to existing theories on stuttering, different predictions about the response patterns could be made. It can on the one hand be expected that people who stutter show greater responses to onset perturbations, since they are supposed to rely to a greater extent on auditory feedback in onsets than fluent speakers (Civier *et al.*, 2010). On the other hand, Harrington (1988) suggests an asynchrony in perceived vowel start vs. expected vowel start in people who stutter, whereby the vowel is perceived earlier than expected. If the onset is stretched again in perturbation (as done in our previous paradigms), the vowel is perceived delayed and could cancel out the perturbatory mismatch effect for people who stutter, leading to less pronounced responses. In any case, the perturbation paradigm could be enhanced by only stretching segments and not compressing others to only elicit compensatory shortening responses. Compensatory shortening necessarily has to be adaptive and could therefore serve to precisely attribute changes in production to the planning level. Lengthening responses, on the other hand, would be indicative of reactive feedback control or delayed auditory feedback effects on control level. Thus, responses to focally applied temporal alterations can shed light on the level where irregularities occur, such as the planning or control level, and inquire further into the functionality of the feedback-feedforward loop in producing fluent speech.

A second focus on population differences arises from the findings in chapter 4. Since it was shown that structural and individual motor stability and individual auditory acuity shape the responses to temporal feedback alterations, it is assumable that groups that have highly trained auditory or motor skills behave differently under temporally perturbed feedback. One such group is highly trained musicians. Playing an instrument at a very high level requires a large set of fine-grained motor actions (that differ depending on the instrument) and exact hearing abilities regarding tone height, quality, and quantity. Testing musicians with the temporal auditory feedback paradigm and the motor and perception test battery could firstly (ideally) support the results from chapter 4 and further indicate whether training in non-speech motor execution and non-speech auditory acuity shapes speech production mechanisms. To date, the investigations on music/musician perturbations are limited and with a strong focus on error rates or disruptive effects in singing or pressing keys on a keyboard under pitch-shifted or delayed auditory feedback (see e.g., Pfordresher and Mantell, 2012). With a greater focus on compensatory/ adaptive effects, investigating responses to temporally perturbed auditory feedback in spoken and sung sequences by a group of musicians and a group of non-musicians could firstly indicate how timing in music stimuli is controlled, and how these two groups differ based on their level of rhythmic ability and auditory acuity.

5.3.3 Phonemic/Lexical Category

A third worthwhile direction for further investigation concerns the effect of the feedback shift on the phonemic category. Many spectral perturbations altered the spectral shape of a sound towards another sound. The temporal alterations in chapters 2 and 3 did not alter the phoneme durations towards another lexical identification. However, spectral perturbation studies showed that phonemic representation plays a significant role in response to the perceived shift (Mitsuya *et al.*, 2011). Thereby, shifts near or crossing a phoneme boundary elicit stronger responses than within-category shifts (Niziolek and Guenther, 2013). Further, a spectral adaptation study by Bourguignon *et al.* (2014) tested the role of lexical identification in feedback-feedforward interaction. They altered the first formant of a vowel in real words or pseudo-words, whereby the shift elicited either a change in lexical category (e.g., from a real word to a pseudo-word) or not. Their subjects compensated more to shifts that changed the lexical category to a/another real word rather than to a pseudo-word, indicating sensitivity to the lexical category and further suggesting that feedback and feedforward systems "interact to a certain extent with higher-order lexical information" (Bourguignon *et al.*, 2014). The responses further indicated that Ganong's lexical effect phenomenon, where the perceptual boundary between a real word and a pseudo-word is shifted towards the real word, could be extended to speech production. Applying temporal shifts to two phonemes that differ exclusively in duration could firstly show whether greater responses near boundaries are also found in the temporal domain. Secondly, findings could allow for a more decisive conclusion about the dimension of the temporal representation as phoneme inherent or defined by the realization of the larger prosodic context. Further, changing the lexical identification of a word could give a more thorough insight into how structural information on different hierarchical levels influences the feedback-feedforward interplay in speech production. An experiment that alters exclusively the temporal extent of distinctive stimuli might be hard to design in English. However, in German, the vowels /a/ and /a:/ differ almost exclusively in duration and provide minimal pairs that could neatly be used for boundary-crossing temporal manipulations. An investigation pursuing temporal phonemic contrast realizations during temporally perturbed auditory feedback will follow this thesis to give insight into the role of temporal phoneme boundaries in fluent speech. Summarizing, the perturbation experiments of this thesis and the studies by Bourguignon *et al.* (2014), Niziolek and Guenther (2013), and Mitsuya *et al.* (2011) illustrate that fluent speech carries many principled levels of higher-order organization that indeed influence the mechanisms in speech production. Lexical status, stress pattern, syllable structure, and syllable position experienced only mild attention in perturbation studies to date and are only a few of many influencing factors that deserve more attention in investigations on feedback-feedforward interactivity. To conclude, it is inevitable to move on from perturbing sustained vowels or single syllables and rise to the challenge of examining the feedback-feedforward interaction in more complex speech material, accepting and considering all the quality characteristics of natural fluent speech.

In future explorations on speech timing, particular importance should be given to *time*, *duration*, and *coordination* as different key aspects in speech production. This thesis focused on *duration* and *coordination*. The concept of time has fascinated scientists since human consciousness. In human behavior, time is ubiquitous and cannot be left aside as the underlying dimension in modeling any dynamical process. However, as there is no action without time, it remains questionable how explicitly *time* itself needs to be modeled. No human can control time, but rather *duration, coordination* and *timing* with the meaningful placement of information in time. Thereby, time passes, as we speak or stay silent.

"All we have to decide is what to do with the time that is given to us"[3]

[3] (Gandalf in "The Lord of the rings" by J.R.R. Tolkien

Deutsche Zusammenfassung

Diese Dissertation verwendet ein neuartiges Perturbations-Design, um die zeitliche Struktur von gesprochener Sprache und den Einfluss des eigenen auditorischen Feedbacks auf die Echtzeit-Kontrolle und Verformbarkeit temporaler Eigenschaften zu untersuchen.

Weiterhin werden individuelle Kompetenzen in der Perzeption sowie individuelle motorisch-rhythmische Fähigkeiten mit der Verformbarkeit temporaler Eigenschaften von Sprache durch manipuliertes auditorisches Feedback in Verbindung gebracht.

Dazu werden drei Hauptexperimente durchgeführt, deren Ausarbeitung, die Analyse und Ergebnisse in den Kapiteln 2-4 dargestellt und diskutiert werden.

Bisher wurden mit Manipulationen des auditorischen Feedbacks beim Sprechen vor allem spektrale Eigenschaften der Sprache manipuliert, wie zum Beispiel Formant-Frequenzen. Diese Manipulationen führen dann dazu, dass ein Vokal im auditorischen Feedback wie ein anderer Vokal klingt, was zu einer gegensteuernden Reaktion der Sprecher führt, der sogenannten Kompensation. Wenn diese gegensteuernde Produktion über den Zeitraum der Manipulation hinaus anhält, spricht man von Adaptation (siehe Caudrelier and Rochet-Capellan, 2019 für einen Überlick).

Zur spektralen Perturbation wurde in den letzten Jahrzehnten viel geforscht, viel weniger ist allerdings darüber bekannt, ob Sprecher ebenfalls kompensieren oder ihre Repräsentationen aktualisieren (adaptieren) wenn temporale Eigenschaften der Sprache manipuliert werden.

In der Erforschung der Sprachproduktion hat sich gezeigt, dass die temporale Entfaltung einzelner Gesten, die zu bestimmten Sprachlauten führen verschiedene zeitliche Koordinationsmuster aufweisen, je nachdem in welcher Position innerhalb einer Silbe sie realisiert werden. Im Onset einer Silbe überlappen die Gesten zweier Konsonanten beispielsweise mehr als in der Coda einer Silbe. Zudem sind in komplexen Onsets die Konsonanten mit dem folgenden Vokal getimed, was zu einem globalen Koordinationsmuster von Onset und Vokal führt, wobei hingegen Coda Konsonanten

sequenziell jeweils lokal mit dem vorhergehenden Konsonanten bzw. Vokal gekoppelt sind (Browman and Goldstein, 1988).
Aber nicht nur die Position innerhalb der Silbe determiniert die zeitliche Struktur der Sprachentfaltung, sondern auch andere prosodische Muster, wie zum Beispiel Wortbetonung. Vokale in einer betonten Silbe sind länger als die gleichen Vokale in unbetonter Position.

In dieser Dissertation wird die temporale Struktur von Silbenposition (Kapitel 2) und Wortbetonung (Kapitel 3) und deren Kontrolle durch das auditorische Feedback untersucht. Durch die zeitliche Manipulierung des auditorischen Feedbacks beim Sprechen wird die Stabilität der temporalen Strukturen geprüft und die Interaktion von Feedback und Feedforward Mechanismen in der temporalen Entfaltung von flüssiger Sprache. In Kapitel 3 werden dann die Reaktionen zu der Perturbation der Silbenstruktur aus Kapitel 2 in den Zusammenhang mit generellen perzeptiven temporalen Diskriminationsfähigkeiten und rhythmisch-motorischen Fähigkeiten der Probanden in Verbindung gebracht. Diese Untersuchung liefert einen ganzheitlichen Eindruck dazu inwieweit individuelle Fähigkeiten in Feedback- und Feedforward-Mechanismen die temporale Umsetzung flüssiger Sprache beeinflussen.

In Kapitel 2 wurden zur Erforschung der Stabilität der prosodischen Silbenstruktur die zwei Wörter „Pfannkuchen“ und Napfkuchen“ manipuliert. In „Pfannkuchen“ wurde dabei der Onset der ersten Silbe /pf/ gedehnt im auditorischen Feedback und der folgendende Vokal /a/ gestaucht (Onset condition). In „Napfkuchen“ hingegen wurde der Vokal der ersten Silbe /a/ gedehnt und die Coda /pf/ gestaucht (Coda condition). Daten von 33/34 Probanden haben gezeigt, dass mehr für die Perturbation von Vokal+Coda kompensiert wird als für die Perturbation von Onset+Vokal. Dabei wurden für beide Segmente der Coda condition sowie für den Vokal der Onset condition kompensiert, also der Perturbation entgegengesteuert. Keine Reaktion wurde für die Manipulation des Onsets beobachtet. Beide Segmente der Coda condition zeigten außerdem adaptives Verhalten mit kompensatorischer Produktion die über die Perturbationsphasen hinaus anhielt. Dieses Experiment hat gezeigt, dass Sprecher/innen in der Lage sind für zeitliche Perturbationen zu kompensieren und geplante Dauern auch aktualisiert werden können im Feedforward System, jedoch mit Abhängigkeit von der Silbenstruktur. Vokale und Codas sind leichter anzupassen als Onsets. Silbenonsets zeigen eine größere artikulatorische Stabilität verankert im Feedforward System und sind deshalb nicht so leicht beeinflussbar durch Veränderungen im auditorischen Feedback.

Im dritten Kapitel wurden Manipulationen im Stile der Onset condition aus Kapitel 2 auf eine betonte und eine unbetonte Silbe gleicher Struktur innerhalb eines Wortes angewendet. In dem Wort „Tschetschenen" wurde zum einen die erste Silbe manipuliert, indem der Onset /tʃ/ gedehnt wurde und der unbetonte Vokal /e/ gestaucht (Unstressed condition). Zum anderen wurde in der zweiten, betonten Silbe ebenfalls /tʃ/ gedehnt und der betonte Vokal /e/ gestaucht (Stressed condition). Analysen wurden auf verschiedenen Ebenen durchgeführt und haben gezeigt, dass die Manipulation der betonten Silbe (Stressed condition) etwas stärkere kompensatorische Reaktionen hervorruft als in der unbetonten Silbe, insbesondere im Vokal, da das Betonungsmuster durch die Perturbation abgeschwächt wurde. Weitere Analysen auf Phonem- und Wortebene haben gezeigt, dass verschiedene lokale motorische Änderungen (Änderungen in der absoluten Dauer von Lauten) zu verschiedenen globalen Änderungen des Timings (in Wort-normalisierten Dauerverhältnissen) führen. Bei Betrachtung der absoluten Dauern (in ms) konnten die Ergebnisse aus Kapitel 2 der Onset-Stabilität reproduziert werden, da keine Veränderung der Onsets feststellbar war. In späteren Teilen des Wortes wurden jedoch andere Lautdauern entsprechend verändert, sodass Kompensation für die manipulierten Onsets in beiden Konditionen auf der Wortebene festgestellt werden konnte. Die Ergebnisse haben gezeigt, dass die Dimension der temporalen Repräsentation vielschichtig ist und motorische Repräsentationen gegebenenfalls von auditorischen abweichen, beziehungsweise anders realisiert werden können.

In Kapitel 4 wurden die Reaktionen auf die Perturbation der Silbenstruktur aus Kapitel 2 mit rhythmisch-perzeptiven und -motorischen Kompetenzen in Zusammenhang gebracht. Eine große Testbatterie an Finger-Tapping Aufgaben zu verschiedenen Stimuli hat drei Qualitäten motorisch-rhythmischer Fähigkeiten erfasst: Motor Stability, Synchronization Consistency und Synchronization Accuracy. Weiterhin wurden mithilfe adaptiver staircase Perzeptionstests temporale auditorische Diskriminierungsfähigkeiten erfasst. Die zugrunde liegenden Dimensionen der Daten wurden mit einer principal component Analyse extrahiert um die Anzahl der Prädiktoren zu verringern. Im Anschluss wurden die Kompensationswerte aus Kapitel 2 mit allen principal components des Tapping- und Perzeptionsblocks als Prädiktoren gemodelt. Dabei wurde das bootstrapping Verfahren angewendet und 100 Modelle per Kompensationssegment mit verschiedener Probandenzusammensetzung gerechnet. Die Prädiktoren, die in mehr als 70 der 100 Modelle als relevant erachtet wurden, wurden in das finale Modell miteinbezogen. Die Ergebnisse haben gezeigt, dass sowohl individuelle perzeptive sowie motorisch-rhythmische Fähigkeiten zur kompensatorischen Reaktion beitragen. Dabei wurde eine bessere Perzeption mit mehr Kompensation in der Onset condition von Kapitel 2 assoziiert, wohingegen schlechtere rhythmische Tapping Stabilität mit mehr Kompensation in der Coda condition zusammenhing. Die Perzeptionskomponente zeigte auch einen Einfluss auf die Segmente in der Coda condition, jedoch mit Spielraum für Interpretationen, die bis auf weitere Ergebnisse verschoben werden soll.

Mit den drei Investigationen hat diese Dissertation neben der Betrachtung von auditorischen Fähigkeiten hervorgehoben, dass der Sprachproduktionsprozess aus zwei Teilen besteht, Feedback und Feedforwad Mechanismen. So, wie die perzeptive Fähigkeit individuell variieren kann, so kann auch die Fähigkeit, exakte motorische Bewegungsmuster zu produzieren von individuellen Fähigkeiten abhängen. Sprache ist dabei ein sehr spezifischer Bewegungsprozess, der nicht nur individuellen Fähigkeiten unterliegt, sondern auch strukturellen die durch das phonologische System einer Sprache determiniert sind.

Ferner konnte diese Dissertation zeigen, dass auditorisches Feedback eine Rolle bei der zeitlichen Umsetzung flüssiger Sprache spielt, in Abhängigkeit von prosodischen Phänomenen. In komplexer, zusammenhängender Sprache ist zu vermuten, dass Silbenonsets zu einem größerem Teil aus somatosensorischem Feedback Informationen ziehen, und Silbenvokale und -codas eher auf auditorisches Feedback reagieren. Damit einhergehend ist die temporale Struktur von Onsets vermutlich mehr im Feedforwardsystem verankert und hat eine stärker ausgeprägte motorische Repräsentation, wohingegen die temporale Entfaltung in späteren Teilen einer Silbe eher über das auditorische Feedback reguliert wird und temporale Ziele eher in Verhältnis zur höheren prosodischen Einheit realisiert werden (Silbe/Wort). Da jedoch sowohl bei sensorischen Feedback- wie auch Feedforward-Mechanismen individuelle Fähigkeiten mit hineinspielen – wie diese Dissertation gezeigt hat – bleibt diese Interaktion komplex.

Die Ergebnisse dieser Dissertation sind bislang mit keinem Sprachproduktionsmodell zu erklären, und bedürfen weiterer Forschung zur vollständigen Modellierung gesprochener Sprache. In bisherigen Modellen ist entweder eine Adjustierbarkeit der temporalen Gestenentfaltung über das auditorische Feedback nicht vorgesehen, oder die spezifisch prosodischen temporalen Muster sind nicht im Modell verankert.

In diese Sinne bietet diese Arbeit eine Grundlage um die Modellierung zeitlicher Prozesse in der Sprachentfaltung zu überdenken und weiterzuentwickeln.

References

Arnal, L. H., and Giraud, A.-L. (**2012**). "Cortical oscillations and sensory predictions," Trends in Cognitive Sciences **16**, 390-398.

Astruc, L., and Prieto, P. (**2006**). "Stress and accent: Acoustic correlates of metrical prominence in Catalan," in *ITRW on Experimental Linguistics* (Athens, Greece).

Bakst, S., and Niziolek, C. A. (**2021**). "Effects of syllable stress in adaptation to altered auditory feedback in vowels," J. Acoust. Soc. Am. **149**, 708-719.

Bartoń, K. (**2020**). "Mumin: Multi-model inference," R package version 1.43. Available at https://cran.r-project.org/web/packages/MuMIn/index.html (last viewed: June 14, 2021).

Bates, D., Maechler, M., Biolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effect models using lme4," Journal of Statistical Software **67(1)**, 1-48.

Baum, S. R. (**1996**). "Compensation to articulatory perturbation: Perceptual data," J. Acoust. Soc. Am. **99**, 3791-3794.

Beckman, M. E., and Edwards, J. (**1994**). "Articulatory evidence for differentiating stress categories," in *Phonological structure and phonetic form. Papers in laboratory phonology*, edited by P. A. Keating, and M. E. Beckman (Cambridge University Press), pp. 7-33.

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.

Berens, P. (**2009**). "Circstat: A matlab toolbox for circular statistics," Journal of Statistical Software **31(10)**, 1-21.

Beretta, L., and Santaniello, A. (**2016**). "Nearest neighbor imputation algorithms: A critical evaluation," BMC Med Inform Decis Mak **16(Suppl3)**, 74-74.

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 June 14, 2021)).

Bombien, L., Mooshammer, C., and Hoole, P. (**2013**). "Articulatory coordination in word-initial clusters of German," Journal of Phonetics **41**, 546-561.

Bombien, L., Mooshammer, C., Hoole, P., and Kühnert, B. (**2010**). "Prosodic and segmental effects on EPG contact patterns of word-initial German clusters," Journal of Phonetics **38**, 388-403.

Bombien, L., Winkelmann, R., and Scheffers, M. (**2021**). "Wrassp: An R wrapper to the assp library.," R package version 1.0.1. Available at https://cran.r-project.org/web/packages/wrassp/index.html (last viewed: Sptember 30, 2021).

Bourguignon, N. J., Baum, S. R., and Shiller, D. M. (**2014**). "Lexical-perceptual integration influences sensorimotor adaptation in speech," Frontiers in human neuroscience **8**, 208.

Browman, C. P., and Goldstein, L. (**1989**). "Articulatory gestures as phonological units," Phonology **6**, 201-251.

Browman, C. P., and Goldstein, L. M. (**1988**). "Some notes on syllable structure in articulatory phonology," Phonetica **45**, 140-155.

Browman, C. P., and Goldstein, L. M. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155-180.

Browman, C. P., and Goldstein, L. M. (**2000**). "Competing constraints on intergestural coordination and self-organization of phonological structures," Les Cahiers de l'ICP. Bulletin de la Communication 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," Human Brain Mapping **25**, 105-117.

Brunner, J. (**2008**). "Acoustic compensation and articulo-motor reorganisation in perturbed speech," (Humboldt Universität Berlin).

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., and Perkell, J. (**2011**). "The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation," Journal of Speech, Language, and Hearing Research **54**, 727-739.

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," Journal of Phonetics **24**, 209-244.

Byrd, D., and Choi, S. (**2010**). "At the juncture of prosody, phonology, and phonetics—the interaction of phrasal and syllable structure in shaping the timing of consonant gestures," Laboratory phonology **10**, 31-59.

Byrd, D., and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," Journal of Phonetics **31**, 149-180.

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., Beal, D. S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2014**). "Impaired timing adjustments in response to time-varying auditory perturbation during connected speech production in persons who stutter," Brain and language **129**, 24-29.

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., and Perkell, J. S. (**2012**). "Weak responses to auditory feedback perturbation during articulation in persons who stutter: Evidence for abnormal auditory-motor transformation," PloS one **7**, e41830.

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," The Journal of Neuroscience **31**, 16483-16490.

Campbell, N., and Beckman, M. (**1997**). "Stress, prominence, and spectral tilt," in *Intonation: Theory, Models and Applications (Proceedings of an ESCA Workshop, September 18-20, 1997,* edited by Antonis Botinis, Georgios Kouroupetroglou, and G. Carayiannis (Athens, Greece), pp. 67-70.

Carillo, K., Doutres, O., and Sgard, F. (**2020**). "Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation," J. Acoust. Soc. Am. **147**, 3476-3489.

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, United States), pp. 2095 - 2099.

Caudrelier, T., and Rochet-Capellan, A. (**2019**). "Changes in speech production in response to formant perturbations: An overview of two decades of research," in *Speech production and perception: Learning and memory*, edited by Susanne Fuchs, Joanne Cleland, and A. Rochet-Capellan (Peter Lang, Berlin), pp. 15-75.

Caudrelier, T., Schwartz, J.-L., Perrier, P., Gerber, S., and Rochet-Capellan, A. (**2018**). "Transfer of learning: What does it tell us about speech production units?," Journal of Speech, Language, and Hearing Research **61**, 1613-1625.

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**, EL355-360.

Cheng, H.-S., Niziolek, C. A., Buchwald, A., and McAllister, T. (**2021**). "Examining the relationship between speech perception, production distinctness, and production variability," Frontiers in Human Neuroscience **15**, 660948.

Cheveigné, A. d., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917-1930.

Cho, T. (**2009**). "Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English," in *Laboratory phonology 8* (De Gruyter Mouton), pp. 519-548.

Cho, T., and Keating, P. (**2009**). "Effects of initial position versus prominence in English," Journal of Phonetics **37**, 466-485.

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," Journal of Fluency Disorders **35**, 246-279.

Cummins, F., and Port, R. (**1998**). "Rhythmic constraints on stress timing in English," Journal of Phonetics **26**, 145-171.

Dalla Bella, S., Farrugia, N., Benoit, C.-E., Begel, V., Verga, L., Harding, E., and Kotz, S. A. (**2017**). "BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities," Behavior research methods **49**, 1128-1145.

Damian, M. F., and Dumay, N. (**2007**). "Time pressure and phonological advance planning in spoken production," Journal of Memory and Language **57**, 195-209.

De Jong, K. J. (**1995**). "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," J. Acoust. Soc. Am. **97**, 491-504.

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," Human Movement Science **31**, 801-810.

Donath, T. M., Natke, U., and Kalveram, K. T. (**2002**). "Effects of frequency-shifted auditory feedback on voice F0 contours in syllables," J. Acoust. Soc. Am. **111**, 357-366.

Drake, C., Jones, M. R., and Baruch, C. (**2000**). "The development of rhythmic attending in auditory sequences: Attunement, referent period, focal attending," Cognition **77**, 251-288.

Efron, B. (**1992**). "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in statistics: Methodology and distribution*, edited by S. Kotz, and N. L. Johnson (Springer, New York, NY), pp. 569-593.

El Zarka, D., Schuppler, B., Lozo, C., Eibler, W., and Wurzwallner, P. (**2015**). "Acoustic correlates of stress and accent in Standard Austrian German," in *Phonetik in und über Österreich, Veröffentlichung zur Linguistik und Kommunikationsforschung*, edited by Sylvia Moosmüller, Carolin Schmid, and M. Sellner (Verlag der Österreichischen Akademie der Wissenschaften), pp. 15-44.

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," Frontiers in Human Neuroscience **8:467**, 1-10.

Falk, S., Müller, T., and Dalla Bella, S. (**2015**). "Non-verbal sensorimotor timing deficits in children and adolescents who stutter," Frontiers in Psychology **6:847**, 1-12.

Fisher, N. I. (**1995**). *Statistical analysis of circular data* (Cambridge university press).

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., and Poeppel, D. (**2019**). "Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries," Nature Human Behaviour **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," Nature Communications **11:2839**, 1-12.

Folkins, J. W., and Zimmermann, G. N. (**1982**). "Lip and jaw interaction during speech: Responses to perturbation of lower-lip movement prior to bilabial closure," J. Acoust. Soc. Am. **71**, 1225-1233.

Fowler, C. A. (**1981a**). "Production and perception of coarticulation among stressed and unstressed vowels," Journal of Speech, Language, and Hearing Research **24**, 127-139.

Fowler, C. A. (**1981b**). "A relationship between coarticulation and compensatory shortening," Phonetica **38**, 35-50.

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., and Hagoort, P. (**2017**). "Individual variability as a window on production-perception interactions in speech motor control," J. Acoust. Soc. Am. **142**, 2007-2018.

Fry, D. B. (**1955**). "Duration and intensity as physical correlates of linguistic stress," J. Acoust. Soc. Am. **27**, 765-768.

Fry, D. B. (**1958**). "Experiments in the perception of stress," Language and speech **1(2)**, 126-152.

Gentilucci, M., Santunione, P., Roy, A. C., and Stefanini, S. (**2004**). "Execution and observation of bringing a fruit to the mouth affect syllable pronunciation," European Journal of Neuroscience **19**, 190-202.

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: A Journal of Neurology **120**, 1587-1602.

Gerrits, E., and Schouten, M. (**2004**). "Categorical perception depends on the discrimination task," Perception & psychophysics **66**, 363-376.

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F. H., Lane, H., and Perkell, J. S. (**2010**). "An investigation of the relation between sibilant production and somatosensory and auditory acuity," J. Acoust. Soc. Am. **128**, 3079-3087.

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*, edited by Gunnar Fant, Hiroya Fujisaki, and J. Shen (The Commercial Press, Beijing), pp. 239-249.

Goldstein, L., and Pouplier, M. (**2014**). "The temporal organization of speech," in *The oxford handbook of language production*, edited by Matthew Goldrick, Victor 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 Therapy Perspectives **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," Proceedings of the National Academy of Sciences of the United States of America **107(25)**, 11597-11601.

Guenther, F. H. (**2003**). "Neural control of speech movements," in *Phonetics and phonology in language comprehension and production: Differences and similarities*, edited by Niels Schiller, and A. Meyer (Walter de Gruyter, Berlin), pp. 209-240.

Guenther, F. H. (**2006**). "Cortical interactions underlying the production of speech sounds," Journal of Communication Disorders **39**, 350-365.

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

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain and Language **96**, 280-301.

Guenther, F. H., and Vladusich, T. (**2012**). "A neural theory of speech acquisition and production," Journal of neurolinguistics **25**, 408-422.

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," Experimental Brain Research **130**, 133-141.

Harrington, J. (**1988**). "Stuttering, delayed auditory feedback, and linguistic rhythm," Journal of Speech, Language, and Hearing Research **31**, 36-47.

Hickok, G. (**2012**). "Computational neuroanatomy of speech production," Nature reviews neuroscience **13**, 135-145.

Houde, J., and Nagarajan, S. (**2011**). "Speech production as state feedback control," Frontiers in Human Neuroscience **5:82**, 1-14.

Houde, J. F., and Chang, E. F. (**2015**). "The cortical computations underlying feedback control in vocal production," Current Opinion in Neurobiology **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," Journal of Speech, Language, and Hearing Research **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," Journal of Speech, Language, and Hearing Research **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 of the New York Academy of Sciences **1169**, 58-73.

Jessen, M. (**1993**). "Stress conditions on vowel quality and quantity in German," Working Papers of the Cornell Phonetics Laboratory **8**, 1-27.

Jessen, M., Marasek, K., Schneider, K., and Claßen, K. (**1995**). "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German," in *Proceedings of the International Congress of Phonetic Sciences* (Stockholm University Stockholm), pp. 428-431.

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," Journal of Phonetics **30**, 303-320.

Kaiser, H. F. (**1970**). "A second generation little jiffy," Psychometrika **35**, 401-415.

Kaiser, H. F., and Dickman, K. W. (**1959**). "Analytic determination of common factors," American Psychologist **14**, 425-425.

Karlin, R., Naber, C., and Parrell, B. (**2021**). "Auditory feedback is used for adaptation and compensation in speech timing," Journal of Speech, Language, and Hearing Research.

Katseff, S., Houde, J., and Johnson, K. (**2012**). "Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback?," Language and Speech **55**, 295-308.

Kawamoto, A. H., Liu, Q., Mura, K., and Sanchez, A. (**2008**). "Articulatory preparation in the delayed naming task," Journal of Memory and Language **58**, 347-365.

Kello, C. T., Plaut, D. C., and MacWhinney, B. (**2000**). "The task dependence of staged versus cascaded processing: An empirical and computational study of stroop interference in speech perception," Journal of Experimental Psychology: General **129**, 340-360.

Kelso, J. S., Tuller, B., Vatikiotis-Bateson, E., and Fowler, C. A. (**1984**). "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures," Journal of Experimental Psychology: Human Perception and Performance **10**, 812-832.

Klein, E., Brunner, J., and Hoole, P. (**2019**). "The relevance of auditory feedback for consonant production: The case of fricatives," Journal of Phonetics **77**, 100931.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (**2005**). "Loudness predicts prominence: Fundamental frequency lends little," J. Acoust. Soc. Am. **118**, 1038-1054.

Koopmans-van Beinum, F. J. (**1994**). "What's in a schwa?," Phonetica **51**, 68-79.

Korzyukov, O., Bronder, A., Lee, Y., Patel, S., and Larson, C. R. (**2017**). "Bioelectrical brain effects of one's own voice identification in pitch of voice auditory feedback," Neuropsychologia **101**, 106-114.

Kotz, S. A., and Schwartze, M. (**2010**). "Cortical speech processing unplugged: A timely subcortico-cortical framework," Trends in Cognitive Sciences **14**, 392-399.

Krause, P. A., and Kawamoto, A. H. (**2019**). "Anticipatory mechanisms influence articulation in the form preparation task," Journal of Experimental Psychology: Human Perception and Performance **45(3)**, 319-335.

Kröger, B. J., Kannampuzha, J., and Neuschaefer-Rube, C. (**2009**). "Towards a neurocomputational model of speech production and perception," Speech Communication **51**, 793-809.

Kröger, B. J., Stille, C. M., Blouw, P., Bekolay, T., and Stewart, T. C. (**2020**). "Hierarchical sequencing and feedforward and feedback control mechanisms in speech production: A preliminary approach for modeling normal and disordered speech," Frontiers in Computational Neuroscience **14**, 573554.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," Journal of Statistical 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," Journal of Neuroscience **32**, 9351-9358.

Lametti, D. R., Smith, H. J., Watkins, K. E., and Shiller, D. M. (**2018**). "Robust sensorimotor learning during variable sentence-level speech," Current Biology **28**, 3106-3113.

Lenth, R., Singman, H., Love, J., Buerkner, P., and Herve, M. (**2018**). "Emmeans: Estimated marginal means, aka least-squares means," R package version 1.6.1. Available at: https://cran.r-project.org/package=emmeans (last viewed June 14, 2021).

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.

Mardia, K. V., and Jupp, P. E. (**2009**). *Directional statistics* (John Wiley & Sons, Chichester).

Martin, C. D., Niziolek, C. A., Duñabeitia, J. A., Perez, A., Hernandez, D., Carreiras, M., and Houde, J. F. (**2018**). "Online adaptation to altered auditory feedback is predicted by auditory acuity and not by domain-general executive control resources," Frontiers in Human Neuroscience **12**.

Max, L., and Gracco, V. L. (**2005**). "Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter," Journal of Speech, Language, and Hearing Research **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.

Mooshammer, C., and Geng, C. (**2008**). "Acoustic and articulatory manifestations of vowel reduction in German," Journal of the International Phonetic Association, 117-136.

Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual centers (p-centers)," Psychological Review **83**, 405.

Mücke, D., and Grice, M. (**2014**). "The effect of focus marking on supralaryngeal articulation–is it mediated by accentuation?," Journal of Phonetics **44**, 47-61.

Munhall, K., Fowler, C., Hawkins, S., and Saltzman, E. (**1992**). "“Compensatory shortening” in monosyllables of spoken English," Journal of Phonetics **20**, 225-239.

Munhall, K. G., Löfqvist, A., and Kelso, J. S. (**1994**). "Lip–larynx coordination in speech: Effects of mechanical perturbations to the lower lip," J. Acoust. Soc. Am. **95**, 3605-3616.

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.

Natke, U., and Kalveram, K. T. (**2001**). "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," Journal of Speech, Language, and Hearing Research **44**, 577-584.

Nault, D. R., and Munhall, K. G. (**2020**). "Individual variability in auditory feedback processing: Responses to real-time formant perturbations and their relation to perceptual acuity," J. Acoust. Soc. Am. **148**, 3709-3721.

Niziolek, C. A., and Guenther, F. H. (**2013**). "Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations," The Journal of Neuroscience **33**, 12090-12098.

Niziolek, C. A., Nagarajan, S. S., and Houde, J. F. (**2013**). "What does motor efference copy represent? Evidence from speech production," Journal of Neuroscience **33**, 16110-16116.

Nozaradan, S., Peretz, I., and Mouraux, A. (**2012**). "Selective neuronal entrainment to the beat and meter embedded in a musical rhythm," Journal of Neuroscience **32**, 17572-17581.

Oschkinat, M., and Hoole, P. (**2020**). "Compensation to real-time temporal auditory feedback perturbation depends on syllable position," J. Acoust. Soc. Am. **148**, 1478-1495.

Oschkinat, M., and Hoole, P. (**2022**). "Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables," Journal of Phonetics **91**, 101133.

Oschkinat, M., Hoole, P., Falk, S., and Dalla Bella, S. (**2022**). "Temporal malleability to auditory feedback perturbation is modulated by rhythmic abilities and auditory acuity," Frontiers in Human Neuroscience **16**.

Parrell, B., Goldstein, L., Lee, S., and Byrd, D. (**2014**). "Spatiotemporal coupling between speech and manual motor actions," Journal of phonetics **42**, 1-11.

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., and Niziolek, C. A. (**2021**). "Increased speech contrast induced by sensorimotor adaptation to a nonuniform auditory perturbation," Journal of Neurophysiology **125**, 638-647.

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 computational Biology **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," Journal of Speech, Language, and Hearing Research **54**, 1051-1059.

Patel, R., Reilly, K. J., Archibald, E., Cai, S., and Guenther, F. H. (**2015**). "Responses to intensity-shifted auditory feedback during running speech," Journal of Speech, Language, and Hearing Research **58**, 1687-1694.

Patri, J.-F., Diard, J., and Perrier, P. (**2019**). "Modeling sensory preference in speech motor planning: A Bayesian modeling framework," Frontiers in Psychology **10:2339**, 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 Computational Biology **14**, e1005942.

Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Frontiers in Psychology **3:320**, 1-17.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., and Zandipour, M. (**2004a**). "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," J. Acoust. Soc. Am. **116**, 2338-2344.

Perkell, J. S., Lane, H., Ghosh, S., Matthies, M. L., Tiede, M., Guenther, F., and Ménard, L. (**2008**). "Mechanisms of vowel production: Auditory goals and speaker acuity," in *Proceedings of the 8th International Seminar on speech production* (Strasbourg, France), pp. 29-32.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., and Guenther, F. H. (**2004b**). "The distinctness of speakers' /s/ - /ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect," Journal of Speech, Language, and Hearing Research **47**, 1259-1269.

Pfordresher, P. Q., and Mantell, J. T. (**2012**). "Effects of altered auditory feedback across effector systems: Production of melodies by keyboard and singing," Acta psychologica **139**, 166-177.

Pouplier, M. (**2012**). "The gestural approach to syllable structure: Universal, language-and cluster-specific aspects," in *Speech planning and dynamics*, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Peter Lang, Frankfurt), pp. 63-96.

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.

Reinisch, E., Jesse, A., and McQueen, J. M. (**2010**). "Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately," Quarterly Journal of Experimental Psychology **63**, 772-783.

Reinisch, E., Jesse, A., and McQueen, J. M. (**2011**). "Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue," Language and Speech **54**, 147-165.

Reinisch, E., and Sjerps, M. J. (**2013**). "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," Journal of Phonetics **41**, 101-116.

Repp, B. H., and Su, Y.-H. (**2013**). "Sensorimotor synchronization: A review of recent research (2006–2012)," Psychonomic Bulletin & Review **20**, 403-452.

Rochet-Capellan, A., Laboissière, R., Galván, A., and Schwartz, J.-L. (**2008**). "The speech focus position effect on jaw-finger coordination in a pointing task," Journal of Speech, Language, and Hearing Research **51**, 1507-1521.

Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., and ten Bosch, L. (**2020**). "Control of speaking rate is achieved by switching between qualitatively distinct cognitive "gaits": Evidence from simulation," Psychological Review **127**, 281-304.

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," Journal of Neurophysiology **43**, 118-136.

RStudio, T. (**2015**). "Rstudio: Integrated development for R " (Inc. RStudio, Boston, MA).

Saltzman, E., and Byrd, D. (**2000**). "Task-dynamics of gestural timing: Phase windows and multifrequency rhythms," Human Movement Science **19**, 499-526.

Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L. (**2008**). "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008), Campinas, Brazil*, pp. 175-184.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological Psychology **1**, 333-382.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (**2016**). "Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models," The R journal **8**, 289-317.

Shaw, J. A., and Chen, W.-r. (**2019**). "Spatially conditioned speech timing: Evidence and implications," Frontiers in Psychology **10:2726**, 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.

Sluijter, A. M. C., and van Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress," J. Acoust. Soc. Am. **100**, 2471-2485.

Sluijter, A. M. C., van Heuven, V. J., and Pacilly, J. J. A. (**1997**). "Spectral balance as a cue in the perception of linguistic stress," J. Acoust. Soc. Am. **101**, 503-513.

Sóskuthy, M. (**2017**). "Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction," arXiv preprint arXiv:1703.05339.

Sowiński, J., and Dalla Bella, S. (**2013**). "Poor synchronization to the beat may result from deficient auditory-motor mapping," Neuropsychologia **51**, 1952-1963.

Subramaniam, K., Kothare, H., Mizuiri, D., Nagarajan, S. S., and Houde, J. F. (**2018**). "Reality monitoring and feedback control of speech production are related through self-agency," Frontiers in human neuroscience **12:82**, 1-8.

Suomi, K., Toivanen, J., and Ylitalo, R. (**2003**). "Durational and tonal correlates of accent in finnish," Journal of Phonetics **31**, 113-138.

Suomi, K., and Ylitalo, R. (**2004**). "On durational correlates of word stress in finnish," Journal of Phonetics **32**, 35-63.

Teki, S., Grube, M., and Griffiths, T. D. (**2012**). "A unified model of time perception accounts for duration-based and beat-based timing mechanisms," Frontiers in integrative neuroscience **5:90**, 1-7.

Teki, S., Grube, M., Kumar, S., and Griffiths, T. D. (**2011**). "Distinct neural substrates of duration-based and beat-based auditory timing," Journal of Neuroscience **31**, 3805-3812.

The MathWorks Inc. (**2012a**). "Matlab [computer program]."

Tierney, A., and Kraus, N. (**2014**). "Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (path)," Frontiers in Human Neuroscience **8:949**, 1-9.

Tilsen, S. (**2016**). "Selection and coordination: The articulatory basis for the emergence of phonological structure," Journal of Phonetics **55**, 53-77.

Tilsen, S. (**2019**). "Space and time in models of speech rhythm," Annals of the New York Academy of Sciences **1453**, 47-66.

Tourville, J. A., Cai, S., and Guenther, F. (**2013**). "Exploring auditory-motor interactions in normal and disordered speech," Proceedings of Meetings on Acoustics **19, 060180**, 1-8.

Tourville, J. A., and Guenther, F. H. (**2011**). "The DIVA model: A neural theory of speech acquisition and production," Language and Cognitive 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.

Turk, A. E., and Sawusch, J. R. (**1996**). "The processing of duration and intensity cues to prominence," J. Acoust. Soc. Am. **99**, 3782-3790.

Turk, A. E., and Shattuck-Hufnagel, S. (**2014**). "Timing in talking: What is it used for, and how is it controlled?," Philosophical Transactions of the Royal Society B: Biological Sciences **369:20130395**, 1-13.

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, (last viewed June 14 2021).

Villacorta, V., Perkell, J., and Guenther, F. (**2005**). "Relations between speech sensorimotor adaptation and perceptual acuity," J. Acoust. Soc. Am. **117**, 2618-2619.

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.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (**2019**). "Welcome to the tidyverse," Journal of Open Source Software **4(43):1686**, 1-6.

Wiese, R. (**2000**). *The phonology of German* (Oxford University Press).

Winkelmann, R., Jänsch, K., Cassidy, S., and Harrington, J. (**2020**). "emuR: Main package of the EMU speech database management system," R package version 2.1.1. Available at https://cran.r-project.org/web/packages/emuR/index.html, (last viewed June 14, 2021).

Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**, 3-36.

Wood, S. N. (**2017**). *Generalized additive models: An introduction with R* (Chapman and Hall/CRC, 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.

Yates, A. J. (**1963**). "Delayed auditory feedback," Psychological bulletin **60(3)**, 213-232.

Acknowledgements

This thesis is the accumulation of so many wonderful people supporting, motivating, and inspiring me over the last decade. Firstly, I am incredibly grateful to my family for always supporting me, no matter what I planned to do with my life. Many thanks to Adrian Simpson for getting me enthusiastic about phonetics the minute I started my studies in Jena. There were no other courses I voluntarily visited at 8 in the morning with nothing but excitement (and a big mug of coffee). Thank you so much, Melanie Weirich, for everything you taught me beyond annotation, your positive vibes, and your support and advice on how to survive big conferences and the scientific world in general. Also, thanks to you and Stef Jannedy for letting me experience how much fun the supremely serious procedure of doing science can be. A special thanks to Jana Brunner and Eugen Klein for adopting me at HU Berlin for my Master's thesis and for preparing me unwittingly for the PhD position I pursued afterward. I would also like to thank Simone Falk and Simone Dalla Bella for the fruitful cooperation and the warm welcome to Montréal.

Moving on to Munich, my gratitude goes to the whole IPS, especially to Jonathan Harrington, Christopher Carignan, Eva Reinisch, Stefano Coretta, and Michele Gubian for the always constructive feedback in every talk and in the corridors and most valuable statistical advice. Thanks also to Klaus Jänsch, without whom we all would have many more problems. A big thank you goes to to Markus Jochim, Ulrich Reubold, and Raphael Winkelmann for teaching me everything I know in R or, let's say, until I could find my own solutions to my own problems. I had the great pleasure of sharing an office and many different stages of emotions with my best friends. Thank you, Katrin Wolfswinkler and Markus Jochim for always being there, and Katrin and Raphael for all the productive and procrastinative coffees we had, which kept me sane whenever it became difficult. Many thanks also to Charlie Wiltshire for proofreading and, together with Mona Franke, for supporting me on the final stretch. I would also like to thank all the students who helped with testing and annotating the data; your work was truly indispensable.

Besides the university environment, a huge hug goes to my beloved Abaco-Orchester and my soccer team ESV München-Freimann, which both undoubtedly sustained my psychological and physiological well-being during my PhD.

At the very end of this story, words cannot express my gratitude to you, Phil. I am beyond grateful for everything we shared in the last 4 years. Thank you for choosing me for your project, for all the fruitful discussions, for your always open door and ears, for letting me travel so much, for your patience, for cricket and the music. I couldn't have hoped for a better constellation for my PhD, and it has been primarily due to your encouragement and trust in me that the last few years became a delightful and fulfilling experience for me.