Journal of Phonetics 91 (2022) 101133

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

# Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables

Miriam Oschkinat *, Philip Hoole

*Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstrasse 3, 80799 Munich, Germany*

ARTICLE INFO

*Article history:*
Received 6 April 2021
Received in revised form 24 December 2021
Accepted 3 January 2022
Available online 2 February 2022

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.


## 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, Freedland, Larson & Hain, 1998; Niziolek & Guenther, 2013; Purcell & Munhall, 2006b; Tourville, Reilly & Guenther, 2008; Xu, Larson, Bauer & Hain, 2004). However, not every online reaction is compensatory. An online response that is not necessarily compensatory in direction, and/or 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 &

* Corresponding author.
*E-mail address:* miriamo@phonetik.uni-muenchen.de (M. Oschkinat).

https://doi.org/10.1016/j.wocn.2022.101133
0095-4470/

Jordan, 1998; Mitsuya, MacDonald, Purcell & Munhall, 2011; Purcell & Munhall, 2006a).

In alterations of formant frequencies, shifts mainly targeted isolated vowels or vowels embedded in monosyllabic real-words (Houde & Jordan, 1998, 2002; Mitsuya et al., 2011; Villacorta, Perkell & Guenther, 2007). Consequently, perturbations of vowels in monosyllabic words give insight into the nucleus' control and representation in stressed syllables. Only a few studies perturbed the center of gravity (COG) of fricatives in monosyllabic words with the finding that speakers also compensate in onset (Shiller, Sato, Gracco & Baum, 2009) and coda (Klein, Brunner & Hoole, 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, Smith, Watkins, and Shiller (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, 1981; 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 disyllabic 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, Niziolek, Reilly & Guenther, 2011) or manipulated the intensity of a stressed syllable bidirectionally (Patel, Reilly, Archibald, Cai & Guenther, 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 (Bombien, Mooshammer & Hoole, 2013; Bombien, Mooshammer, Hoole & Kühnert, 2010; Browman & Goldstein, 2000; Byrd, 1996; Byrd & Choi, 2010; Byrd & Saltzman, 2003; Cho & Keating, 2009; Goldstein, Nam, Saltzman & Chitoran, 2009; Goldstein & Pouplier, 2014; Nam, Goldstein & Saltzman, 2009).

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, MacDonald, and Munhall (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, Fuchs, and Kell (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 & 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 (Browman & Goldstein, 2000; Byrd, 1996; Goldstein & 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, Naber, and Parrell (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 (Karlin et al., 2021; Oschkinat & Hoole, 2020).

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 (e.g., in Catalan: Astruc & Prieto, 2006; in Austrian German: El Zarka, Schuppler, Lozo, Eibler & Wurzwallner, 2015; in German: Jessen, 1993; Jessen, Marasek, Schneider & Claßen, 1995; in English: Kochanski, Grabe, Coleman & Rosner, 2005; in Dutch: Sluijter & van Heuven, 1996; Sluijter, van Heuven & Pacilly, 1997). 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 & 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 & 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 & 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) (e.g., Fry, 1958). Later studies considered pitch markings as a correlate of accent rather than an effect of word stress (e.g., Beckman & Edwards, 1994; Sluijter & 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 & 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 naïvelisteners. 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 & 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 & 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 & 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; Karlin et al., 2021; Oschkinat & Hoole, 2020), 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 & 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 & 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 & 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 & 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 (Guenther, Ghosh, & Tourville, 2006; Tourville & Guenther, 2011).

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 3.1), sound/segment durations on the syllable level relative to the applied perturbation (section 3.2), and sound/segment durations on the word level (normalized by word duration) (section 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 3.4). 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, we do not assess changes in f0 in the perturbed sequences here. 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 & 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.

## 2. Methods

### 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$^{TM}$ 3A in-ear earphones with 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 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, Doutres, and Sgard (2020)). The experiment was conducted in MATLAB (The MathWorks Inc., 2012) using the AUDAPTER software package of Cai, Boucek, Ghosh, Guenther, and Perkell (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, Ghosh, Guenther & Perkell, 2011; Tourville, Cai & Guenther, 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.

### 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 /eː/ 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 Fig. 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 Fig. 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

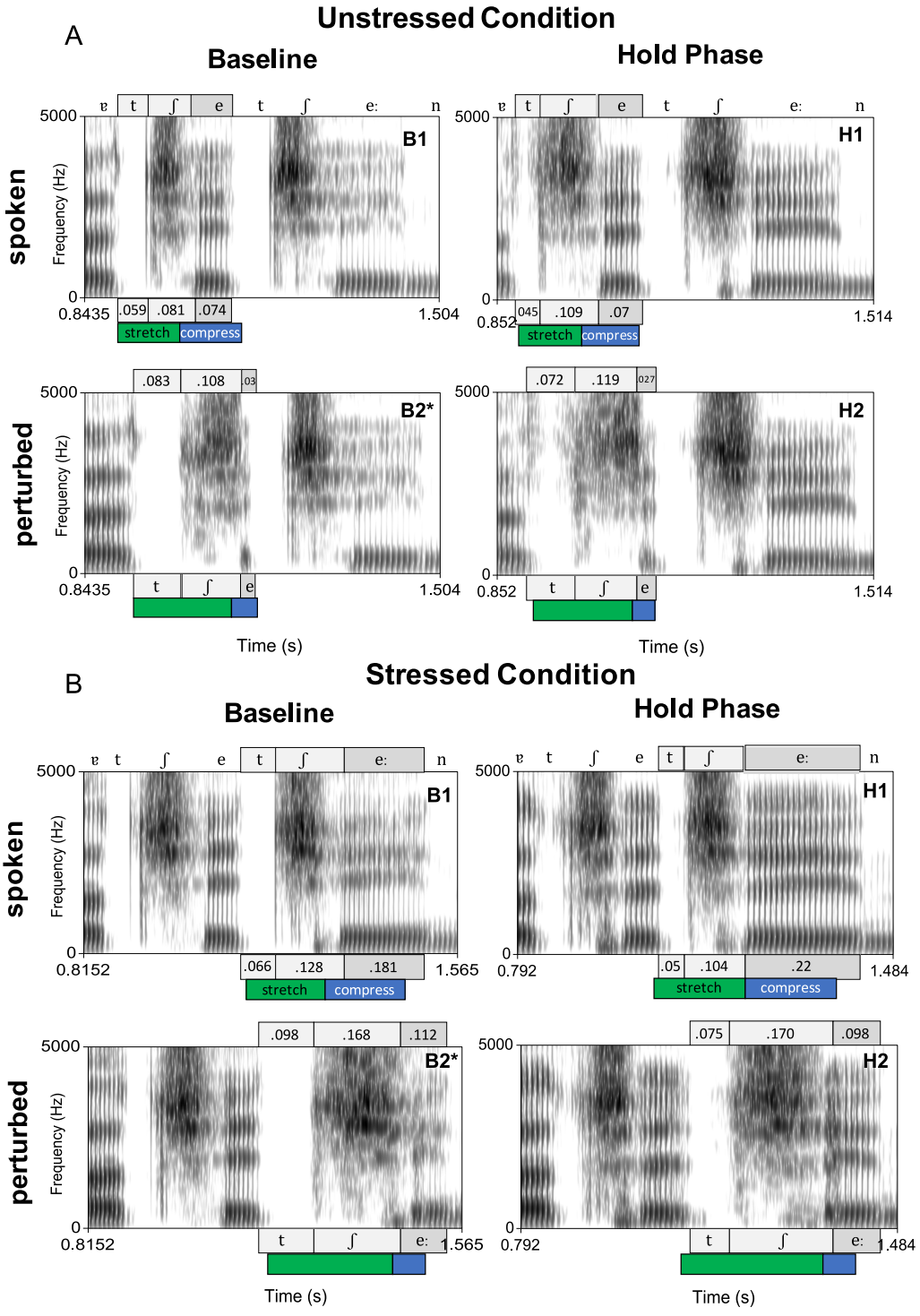



**Fig. 1.** Spectrograms of a baseline and a hold phase trial of a male subject for each condition. The Unstressed condition (perturbation of syllable 1) in the upper plot **(A)**, and the Stressed condition (perturbation of syllable 2) in the lower plot **(B)**. The upper panels per plot show the spoken signal of one baseline trial (**B1**) and a 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),** resepctively. 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*, H2) includes the Audapter delay of 24 ms.

covered by the perturbation section, thus experiencing some shortening (see Figure 2 upper panels).

### 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 (/'tʃeː/), 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.

### 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. 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 3.1). 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 3.2). 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). Sections 3.1 and 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.

### 3.1. Temporal adjustments on the sound/segment level

The first nine baseline trials were discarded from further calculations as done previously (Oschkinat & 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., Fig. 2).

#### 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, Maechler, Biolker & Walker, 2015) and *lmerTest* (v3.1-3; Kuznetsova, Brockhoff & Christensen, 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:

$$formula = duration \sim phase * segment * condition + (phase|Subject), data = dataset_{1/2}.$$

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, Singman, Love, Buerkner & Herve, 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 stan-

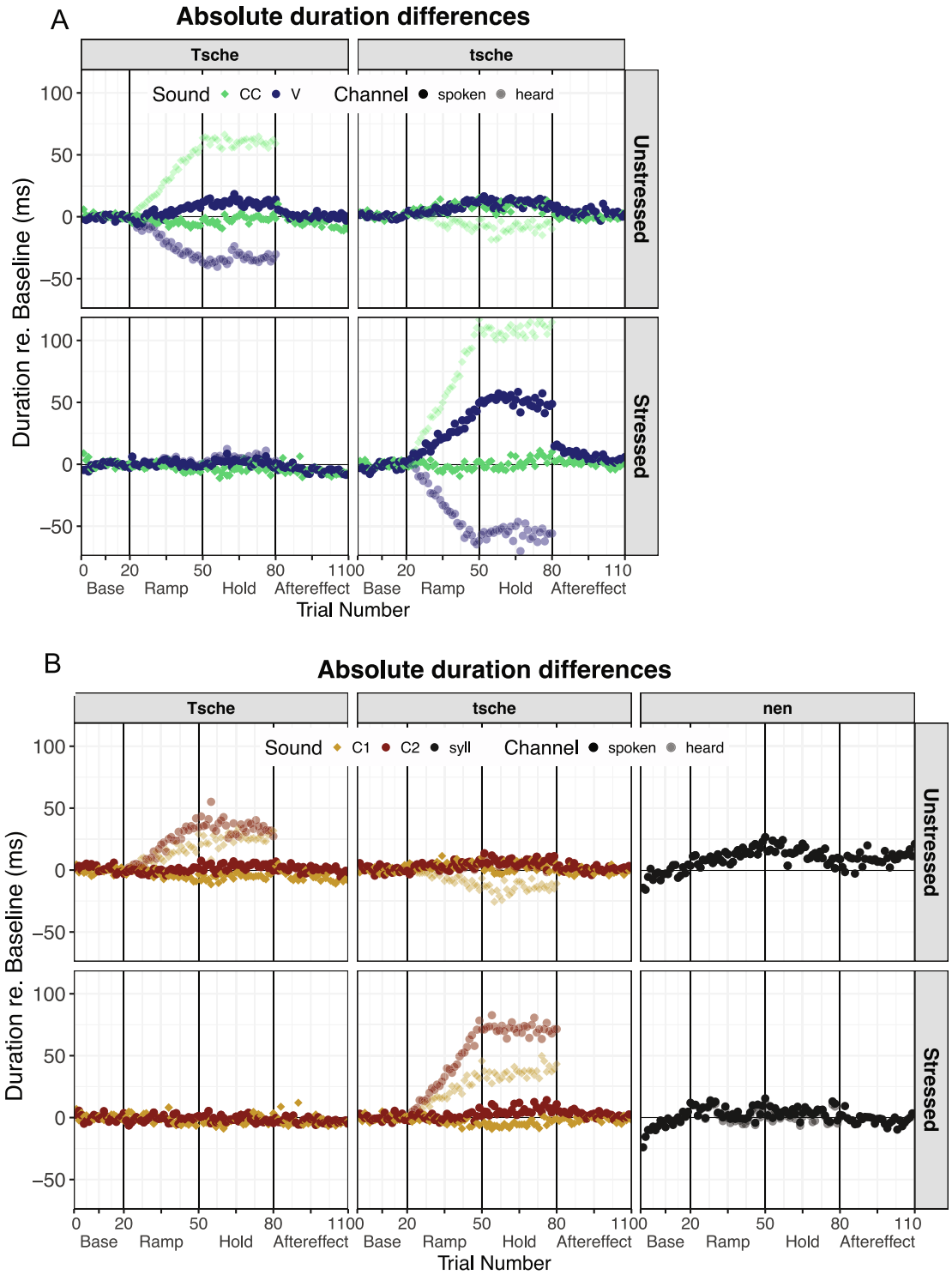



**Fig. 2.** Duration differences in ms relative to the baseline mean of the onset CC (/tʃ/, green) and the vowel (/e/, blue) in syllable 1 and 2 in the upper plot **(A)**, and C1 ([t], orange) and C2 ([ʃ], red) of both syllables as well as syllable 3 (black) in the lower plot **(B)** over the course of the experiment (30 subjects). 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).

dard errors, degrees of freedom, t-ratios, and p-values are presented in Table 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:

**Table 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 |

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%) .

*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 remained. 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, Wieling, Baayen & van Rijn, 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

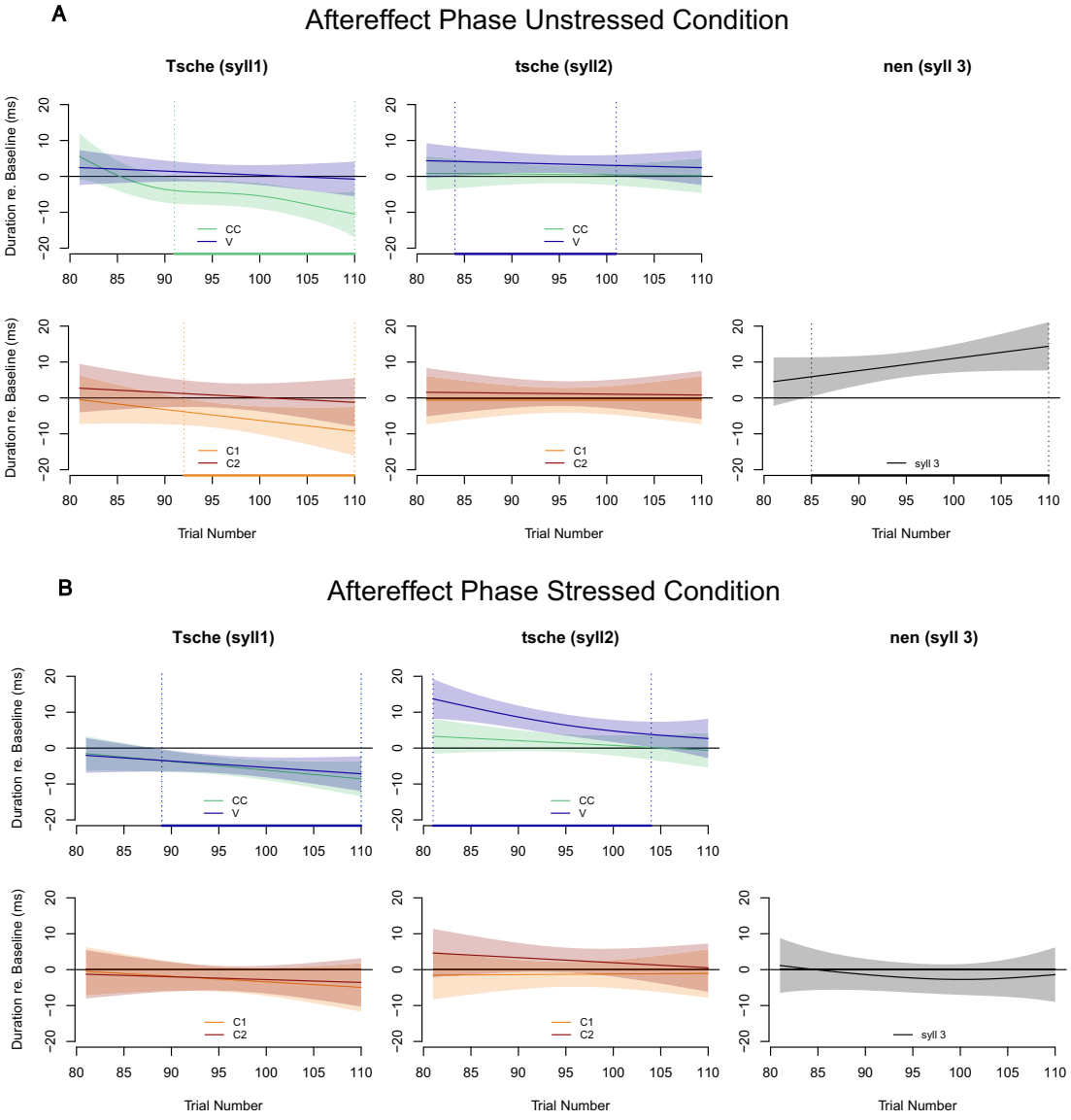



**Fig. 3.** GAMM fits of the aftereffect phase for absolute durations relative to the baseline mean (ms), including random effects and confidence intervals (97.5%). The Unstressed condition visualized in the upper plot **(A)** and the Stressed condition in the lower plot **(B)** (30 subjects). The CC 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.

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 Fig. 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 (Fig. 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 (Fig. 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 (Fig. 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.

### *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)

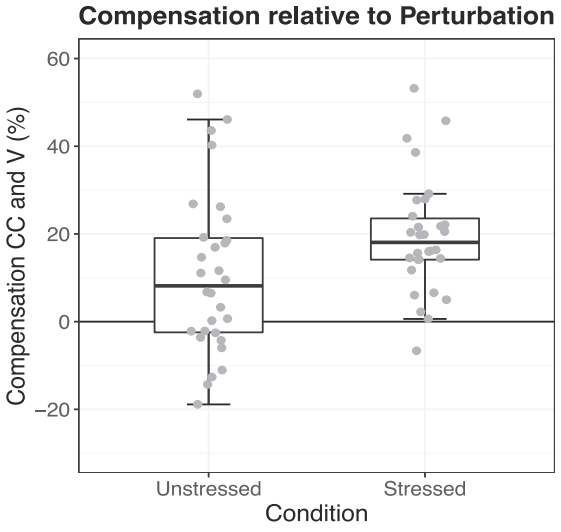



**Fig. 5.** The compensation magnitude relative to perturbation for Unstressed and Stressed condition for 30 subjects. Values incorporate both perturbed segments of interest (CC and V). Boxes correspond to the first and third quartiles and bars represent the median. Whiskers extend from the hinge to the highest/smallest value no further than 1.5 IQR. Data beyond the whiskers are outliers. Dots mark individual subjects.

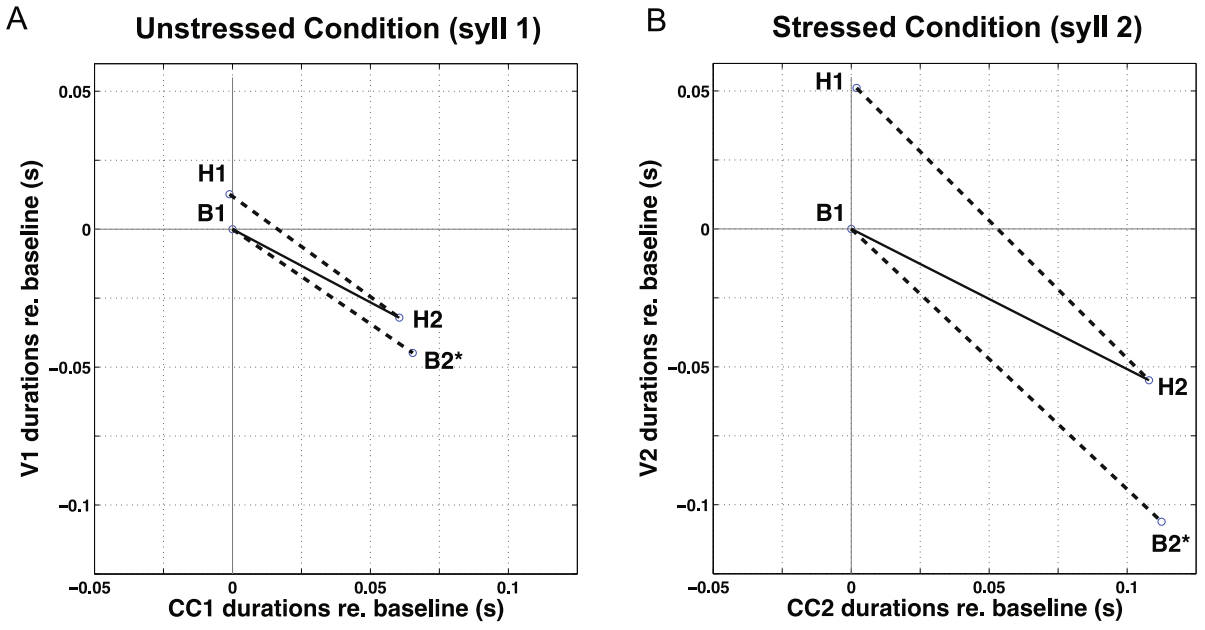



**Fig. 4.** Both plots show mean durations (s) of both segments of interest (CC /tʃ/ and V /e/) over 30 subjects per perturbation condition relative to the baseline mean (0/0). The first segment of the perturbation section is on the x-axis (CC) and the second segment of the perturbation section is on 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.

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 Fig. 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 Figs. 4. A and Figure 4.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*|, Figs. 4.A and Figure 4.B, dashed line) and perturbation on a signal that perhaps already includes a reaction in the hold phase (Euclidian distance |H1-H2|, Figs. 4.A and Figure 4.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.

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

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

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). Fig. 5 visualizes the compensation magnitudes of both conditions.

### 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.

#### 3.3.1. 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 3.1.1 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 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. Fig. 6 depicts the duration per segment within the word (*estimate*) relative to the calculated baseline mean (horizontal zero line) throughout the experiment.

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%).

#### 3.3.2. Adaptation – Evaluation of the aftereffect phase

Similarly to the analyses of absolute durations in section 3.1.2, general additive mixed models (GAMMs) were fitted with the same model structure as described in section 3.1.2, but to

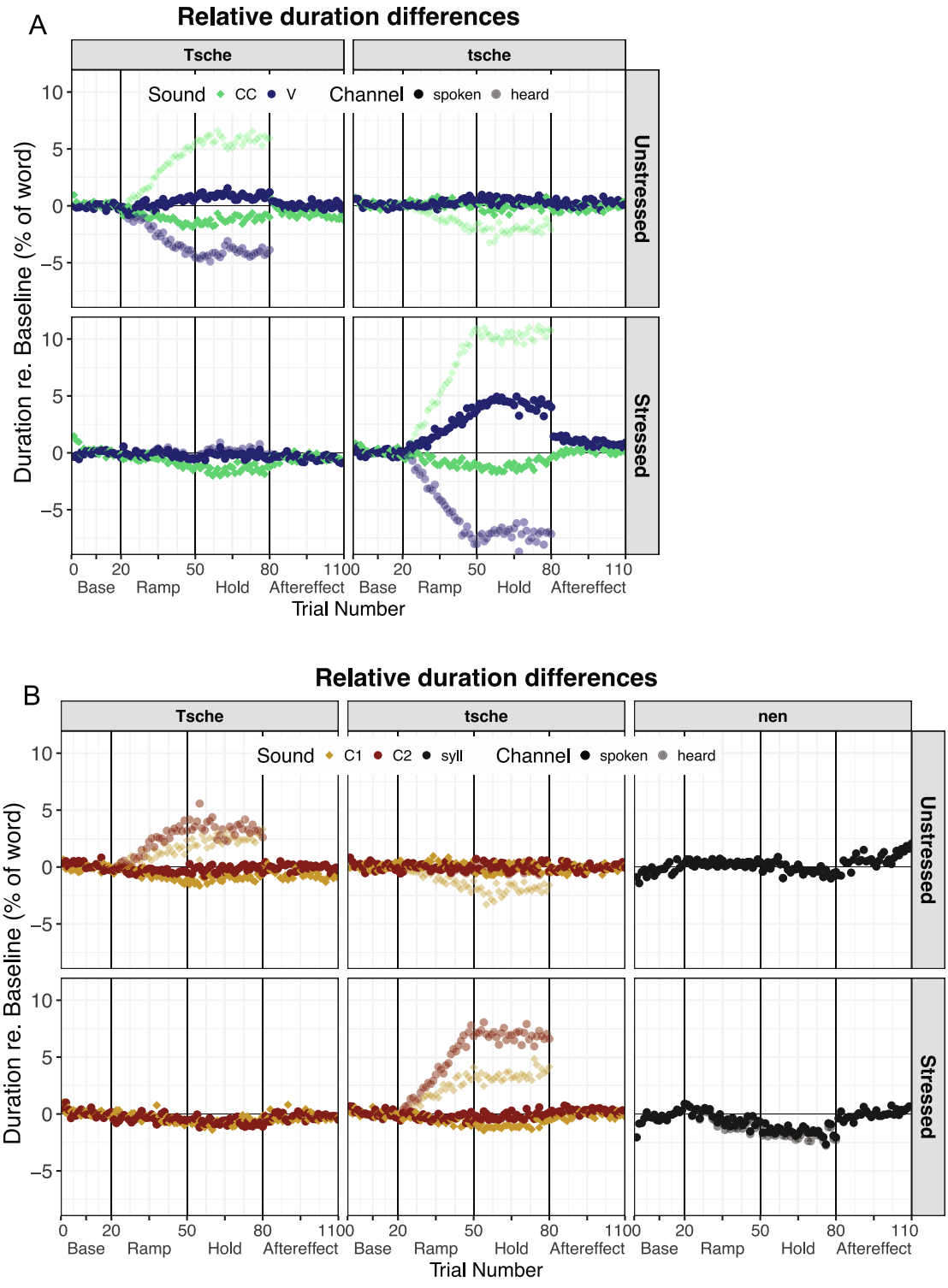



**Fig. 6.** Duration differences relative to the baseline mean (estimate in %) of the onset CC (/tʃ/, green) and the vowel (/e/, blue) in syllable 1 and 2 in the upper plot **(A)**, and C1 ([t], orange) and C2 ([ʃ], red) of both syllables as well as syllable 3 (black) in the lower plot **(B)** over the course of the experiment (30 subjects). 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).

**Table 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** |

*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.

Fig. 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 (Fig. 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.

### 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.

#### 3.4.1. Aperiodicity

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

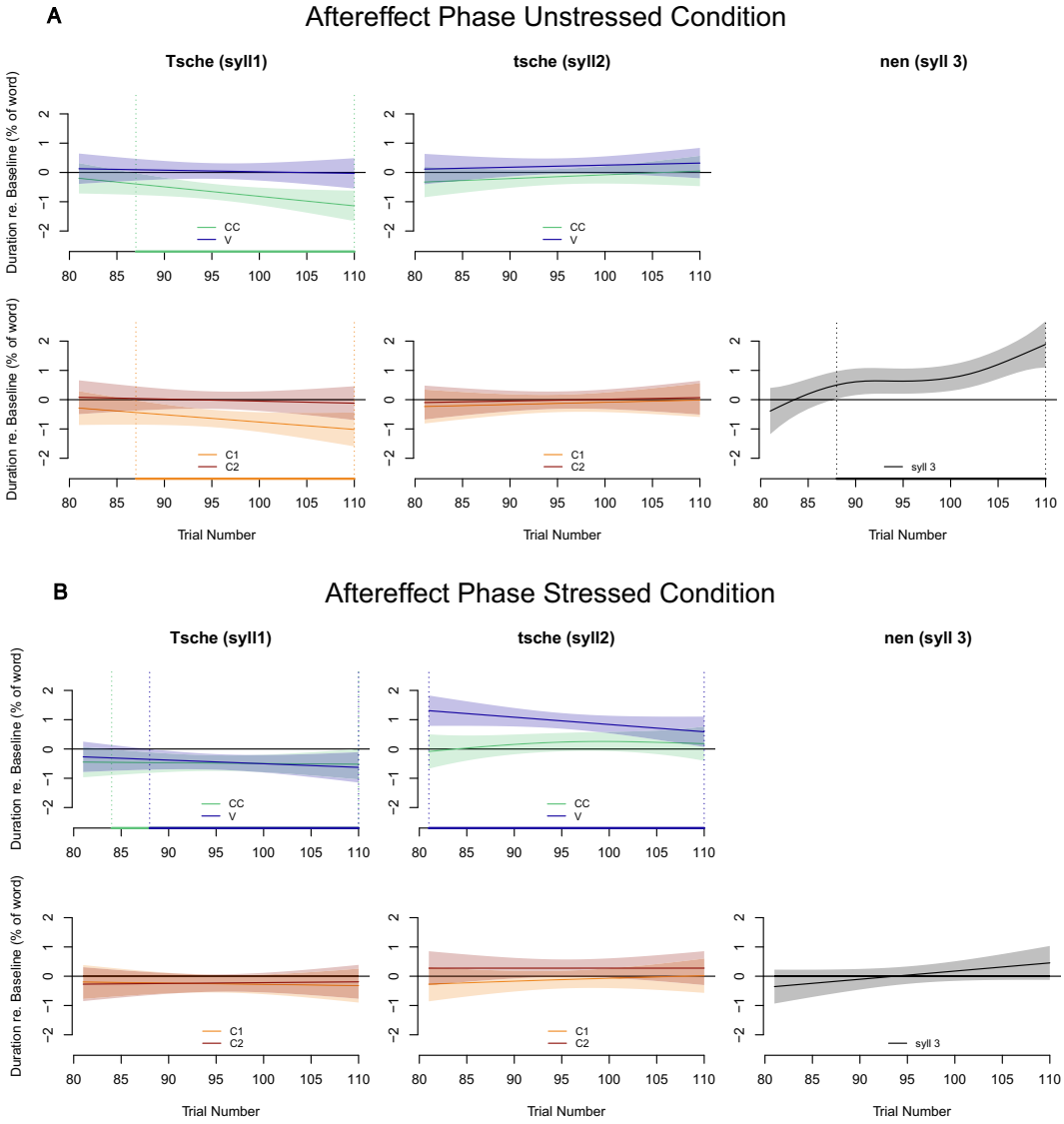



**Fig. 7.** GAMM fits of the aftereffect phase for word normalized durations relative to the baseline mean (%), including random effects and confidence intervals (97.5%). The Unstressed condition in the upper plot **(A)** and the Stressed condition in the lower plot **(B)** (30 subjects). The CC fits 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.

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 Fig. 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; $p < 0.001$). 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; $p < 0.001$). 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; $p < 0.001$).

### *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 subse-

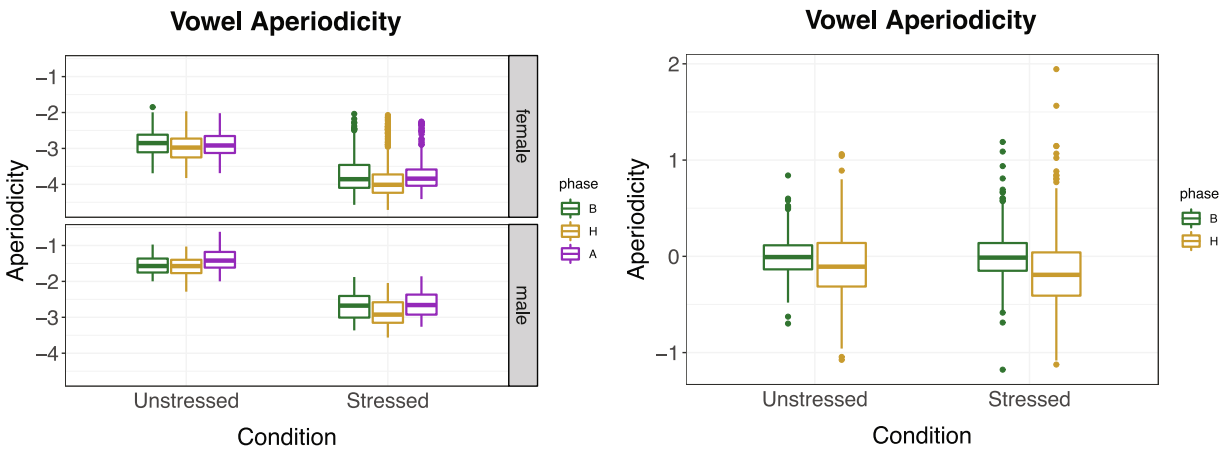



**Fig. 8.** Aperiodicity (log-transformed) of the vowels in both perturbation conditions split by sex (left panel). The right panel shows aperiodicity values (log-transformed) relative to the baseline mean for baseline and hold phase.

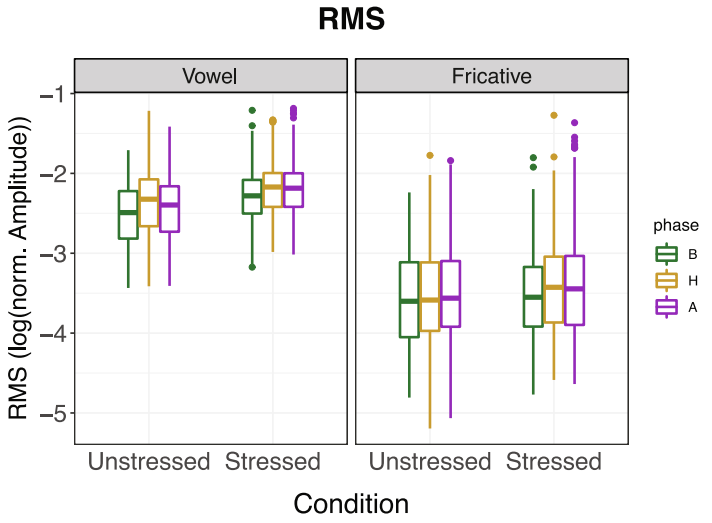



**Fig. 9.** Log-transformed RMS values (y-axis) in both perturbation conditions split by segment (vowel/fricative).

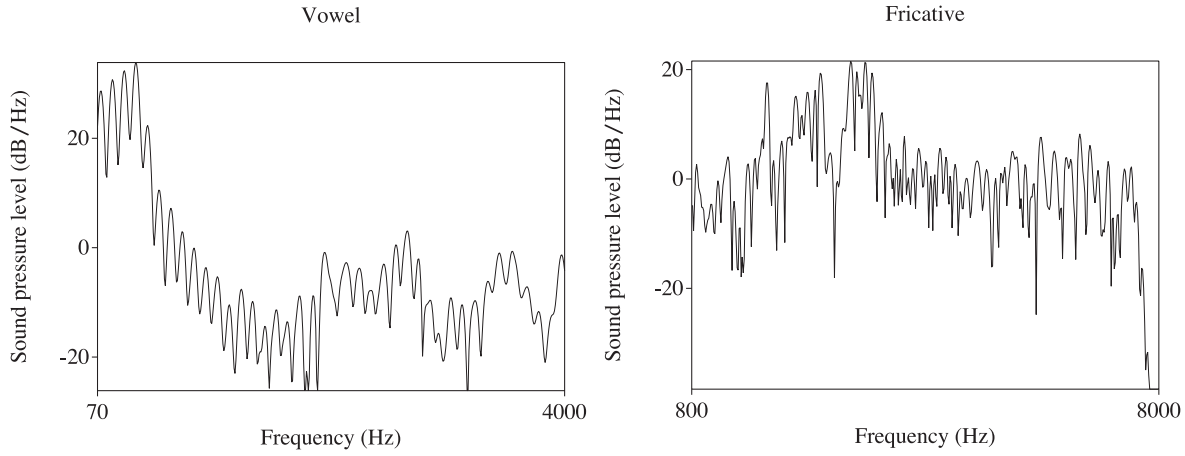



**Fig. 10.** Spectral slices of the vowel /e/ (left panel) and the fricative /ʃ/ (right panel) in the second syllable of the word “Tschetschenen” spoken by a male speaker. Measures were taken in the inner 50% of the sounds in a baseline trial. Both spectra are positively skewed.

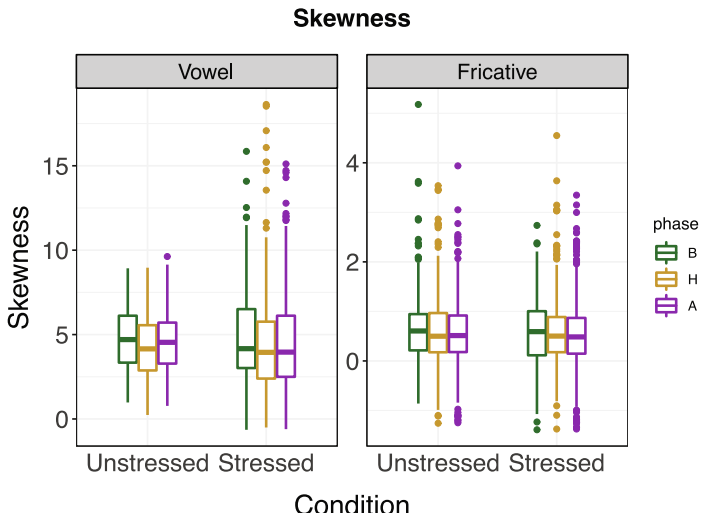



**Fig. 11.** Skewness (y-axis) in both perturbation conditions split by segment (vowel/fricative).

quently 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. Fig. 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; $p < 0.001$) 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; $p < 0.001$, fricative estimate = 0.126; SE = 0.0367; df = 37.8; *t-ratio* = 3.452; $p = 0.0014$).

### 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 (15 ms) for the vowel or 320 samples (20 ms) 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). Fig. 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.

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; $p < 0.001$). 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; $p = 0.67$). Fig. 11 visualizes spectral skewness of the vowel and the fricative in both perturbation conditions.

### 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 frica-

tive /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.

## 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 & 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 3.1. It turns out the patterns found there lead very naturally into a discussion of relative durations at the word level (section 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.

### *4.1. Duration and timing during perturbation*

#### *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 & 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 reorganization 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 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 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 (Cho, 2009; De Jong, 1995; Mücke & 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.

#### 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 Fig. 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 & Choi, 2010).

#### 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 3.2 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 3.2.

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 (Bakst & Niziolek, 2021; Natke & Kalveram, 2001).

### 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 & Hoole, 2020). This interpretation is in line with the perception study by Reinisch, Jesse, and McQueen (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, Jesse & McQueen, 2010; Reinisch et al., 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 Figs. 2 and 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 compensation (reactive feedback control) in the 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, Schwartz, Perrier, Gerber, and Rochet-Capellan (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 & Hoole, 2020).

### 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 aperiodic-

ity 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 & 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.

## 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, for further discussion see Oschkinat and Hoole (2020) and 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 & Shattuck-Hufnagel (2020). While the perturbation of the unstressed word-initial syllable caused a global lengthening of following segments (reactive feedback control), 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 possibilites 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.

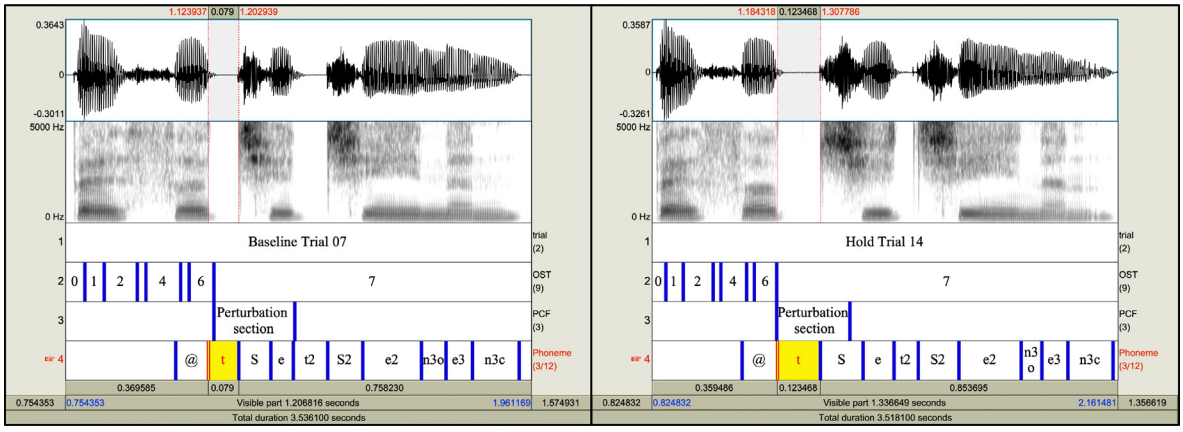
**Acknowledgments**

This work is supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under Grant No. HO 3271/6-1. Thanks to three anonymous reviewers and the associate editor for very constructive comments on a previous version of the manuscript. We also thank Michele Gubian for statistical advice and Sebastian Böhnke, Nicole Benker, Paul Bachmann, and Valeria Trubnikow for their assistance in running the tests.

**Appendix A**

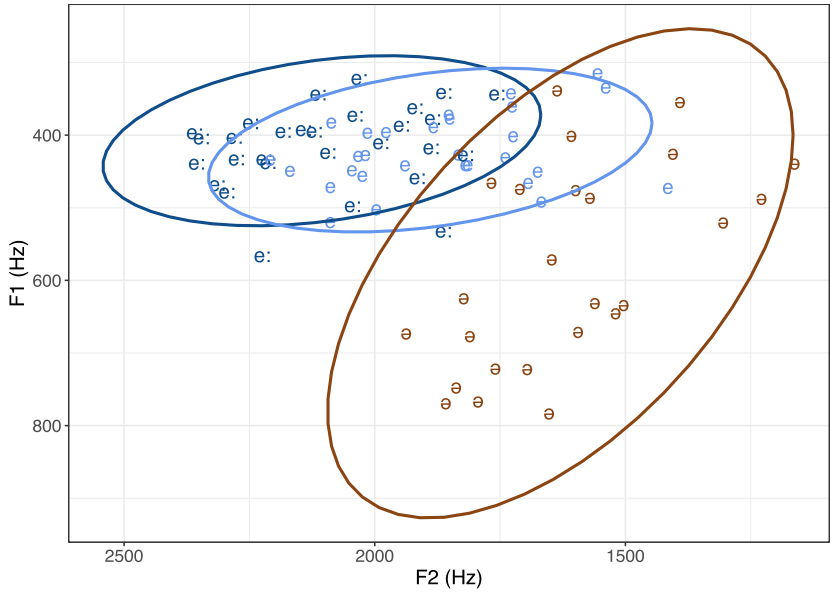
First and second Formants (F1/F2) of the three vowels in “Tschetschenen” (/tʃe'tʃeːnən/). Vowels were provided by the *wrassp* package for signal analysis (Bombien, Winkelmann & Scheffers, 2021) using *EMuR* (Winkelmann, Jänsch, Cassidy & Harrington, 2020). Formants were extracted over all trials of 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.

**Appendix B**




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).

## Appendix C

The following tables report the significance of the interactions received from the linear mixed models calculated in sections 3.1 and 3.3. In model 2 (Table 4), the threeway-interaction was dropped. "Segment" is the concatenation of sound (e.g, CC) and syllable (e.g., syllable 1).

**Table 3**
Statistical outcome of model 1. CC and V with absolute durations (section 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 4**
Statistical outcome of model 2. C1, C2, and syll. 3 with absolute durations (section 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 5**
Statistical outcome of model 3. CC and V with relative durations (section 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 6**
Statistical outcome of model 4. C1, C2, and syll. 3 with relative durations (section 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 |

## References


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

Bakst, S., & Niziolek, C. A. (2021). Effects of syllable stress in adaptation to altered auditory feedback in vowels. *The Journal of the Acoustical Society of America, 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., & Walker, S. (2015). Fitting linear mixed-effect models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

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

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

Bombien, L., Mooshammer, C., Hoole, P., & 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., & 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: September 30, 2021)*.

Browman, C. P., & 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.

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*, 3153–3161.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*, 209–244.

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

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & 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., & 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., & Beckman, M. (1997). Stress, prominence, and spectral tilt. In Antonis Botinis, Georgios Kouroupetroglou & G. Carayiannis (Eds.), *Intonation: Theory, models and applications (Proceedings of an ESCA Workshop, September 18-20, 1997* (pp. 67-70). Athens, Greece.

Carillo, K., Doutres, O., & Sgard, F. (2020). Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America, 147*, 3476–3489.

Caudrelier, T., Schwartz, J.-L., Perrier, P., Gerber, S., & 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.

Cheveigné, A. D., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*, 1917–1930.

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

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

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America, 97*, 491–504.

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

Floegel, M., Fuchs, S., & Kell, C. A. (2020). Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control. *Nature Communications, 11*(2839), 1–12.

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

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*, 765–768.

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

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. In G. Fant, H. Fujisaki, & J. Shen (Eds.), *Frontiers in phonetics and speech science* (pp. 239–249). Beijing: The Commercial Press.

Goldstein, L., & Pouplier, M. (2014). The temporal organization of speech. In M. Goldrick, V. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 210–227). New York: Oxford University Press.

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

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*, 1213–1216.

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research, 45*, 295–310.

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., & 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* (pp. 428–431). Stockholm University Stockholm.

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

Klein, E., Brunner, J., & 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., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*, 1038–1054.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82* (13), 1–26.

Lametti, D. R., Smith, H. J., Watkins, K. E., & 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., & 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)*.

Mitsuya, T., MacDonald, E. N., & Munhall, K. G. (2014). Temporal control and compensation for perturbed voicing feedback. *The Journal of the Acoustical Society of America, 135*, 2986–2994.

Mitsuya, T., MacDonald, E. N., Purcell, D. W., & Munhall, K. G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *The Journal of the Acoustical Society of America, 130*, 2978–2986.

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

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

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupé (Eds.), *Approaches to phonological complexity* (pp. 299–328). Berlin: de Gruyter.

Natke, U., & 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.

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

Oschkinat, M., & Hoole, P. (2020). Compensation to real-time temporal auditory feedback perturbation depends on syllable position. *The Journal of the Acoustical Society of America, 148*, 1478–1495.

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research, 54*, 1051–1059.

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

Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America, 120*, 966–977.

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

Reinisch, E., Jesse, A., & 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., & 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., & 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.

Saltzman, E., Nam, H., Krivokapic, J., & 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).

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America, 125*, 1103–1113.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America, 100*, 2471–2485.

Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America, 101*, 503–513.

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

The MathWorks Inc. (2012a). Matlab [Computer Program]. In.

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

Tourville, J. A., & Guenther, F. A. (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., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*, 1429–1443.

Turk, Alice, & Shattuck-Hufnagel, Stefanie (2020). Timing Evidence for Symbolic Phonological Representations and Phonology-Extrinsic Timing in Speech Production. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02952.

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America, 99*, 3782–3790.

Turk, A. E., & 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., & 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. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America, 122*, 2306–2319.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1–6. 1686.

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

Winkelmann, R., Jänsch, K., Cassidy, S., & 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*. Boca Raton, FL: Chapman and Hall/CRC.

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

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin, 60*(3), 213–232.