The effect of real-time temporal auditory feedback perturbation on the timing of syllable structure

Miriam Oschkinat, Philip Hoole

Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich
Miriam.Oschkinat@phonetik.uni-muenchen.de, hoole@phonetik.uni-muenchen.de

ABSTRACT

Perturbations of auditory feedback (AF) have proven very useful for studying the interaction between feedback and feedforward systems in speech production. AF clearly contributes crucially to planning and execution of spectral speech targets; and subjects typically show compensatory responses in the opposite direction to a feedback manipulation. However, less is known about the reaction to perturbations in the temporal domain and the importance of AF for planning and execution of temporal properties of speech targets. It is not even clear whether compensatory behaviour can in principle occur. Accordingly, this study investigates real-time AF manipulations in the temporal domain, viz. stretching and compressing of absolute durations in onset (CCV) vs. coda (VCC) positions. Since CCV forms a gesturally more cohesive and stable structure than VCC, we expected greater reactions to perturbations in the coda. Compensatory responses were indeed found, and these were overall stronger in the coda.

Keywords: auditory feedback perturbation, temporal domain, syllable structure

1. INTRODUCTION

Over the last couple of decades two approaches to modelling speech production have formed a particular focus of discussion: On the one hand Articulatory Phonology, wherein temporal properties of given gestures and relative timing of active gestures with one another are described, but with no integration of physical auditory or somatosensory feedback and their interaction with a feedforward system [10] (but see [9]). On the other hand, the DIVA model maps cognitive representations to functions of how speech targets are built up, stored and modified through auditory and somatosensory feedback. Then again, DIVA does not incorporate representations of how dynamic specifications of speech emerge and how they are executed [11]. Assuming that speech production is controlled through auditory and somatosensory feedback with the aim of achieving acoustic goals, then speech targets presumably contain information about acoustic features of sounds. One experimental paradigm that supports this idea comes from AF perturbation experiments, in which spectral properties of the AF are altered while the somatosensory feedback remains unchanged. It was demonstrated that subjects react to manipulations of their own AF, e.g. perturbation of formant frequencies, F0, or fricative spectra, mainly with a compensation in the opposite direction to the manipulation [8,6,4]. Driven by the need to combine the modelling of dynamic features described in Articulatory Phonology with the modelled interaction between feedback and feedforward systems in DIVA, this study investigates subjects' reactions to an online AF perturbation in the temporal domain.

In research on prosodic features of speech it was shown that the structure of syllables decisively contributes to speech timing. Within a syllable, onset, nucleus and coda reveal different patterns of temporal flexibility, outlined in two concepts that describe a stronger temporal relationship and close coupling between onset and following vowel compared to vowel and coda in both articulation (c-center effect [3]) and perception (p-center effect [7]).

To get insight into the role of AF for timing mechanisms of syllable structure, manipulations of AF will be applied to onset+vowel (CCV) sequences and vowel+coda (VCC) sequences in a similar phonological and lexical context. The component durations of CCV/VCC sequences will be stretched (first 50% of the sequence) and compressed (second 50% of the sequence) and fed back in real-time to the subject. Subjects' reactions to temporally perturbed AF are expected to give information about the prosodic stability of the syllable structure from an acoustical point of view and the relevance of AF for planning and execution of temporal properties of speech targets.

2. TEMPORAL REAL-TIME MANIPULATION OF AUDITORY FEEDBACK

In spectral (spatial) AF perturbations, predictions of how the intended speech unit should sound do not match the received feedback. In temporal AF perturbations, the time a particular speech event is predicted to happen and its time frame do not match the actual time the speech event is auditorily received.

Cai et al. [1] developed a paradigm that allows fine-grained real-time perturbation of fluent speech utterances. Mainly used for investigations on spectral perturbations, they also conducted a study containing temporal manipulations. In Cai et al. [2] the perceived F2 (second formant) minimum of the vowel [u] in "owe" within the utterance "I owe you a yo-yo" was either accelerated, where the vowel target was perceived earlier in time or decelerated, where the vowel target was perceived later in time. They could show that subjects are sensitive to a perturbation of perceived timing, at least for the deceleration condition with reactions in the same direction as the perturbation (delaying/lengthening of following segments). Their study altered a perceived speech target while keeping the total duration of the speech unit constant. The present study asks how subjects react to a perturbation that does not alter the time point of a speech target, but rather the absolute duration of specific speech sounds and their temporal relationship to each other within a syllable.

Perturbations will be applied to onset and coda segments. Based on the knowledge that onset and vowel form a gesturally more cohesive and stable structure and contribute more to syllable timing than vowel and coda [3], we expect greater reactions to temporal perturbations of VCC (coda) sequences than to perturbations of CCV (onset) sequences.

2.1. Subjects and speech material

For this study 23 participants aged between 19 and 30 (ø23y, 18 females) without any hearing or speech disorders were recruited. All subjects performed both perturbation conditions (onset and coda manipulations). The order of testing was counterbalanced over subjects. For technical reasons, two subjects had to be excluded for the onset condition and four (others) for the coda condition, resulting in a set of 21 subjects for the onset condition and 19 for the coda condition.

For comparable manipulation of onset and coda segments, stimuli with the same sound sequences in different syllable positions were needed. Both words ought to have a similar lexical frequency and phonological surrounding. Therefore, for the onset condition the word "Pfannkuchen" (/ˈpfanku:xən/, *pancake*) was chosen, with perturbation applied to the onset affricate and vowel of the first syllable (/pfa/), while for the coda condition the word "Napfkuchen" (/ˈnapfku:xən/, *ring cake*) was selected, with perturbation of the vowel and coda affricate of the first syllable (/apf/). To allow the software stable real-time tracking for triggering of the intended perturbation section (as described below), the testwords were spoken after the carrier word "besser" (/ˈbɛsɐ/, *better*), resulting in the German phrases "besser Pfannkuchen" or "besser Napfkuchen".

2.2. Experimental setup

Subjects were provided with EAR-Tone in-ear headphones for perturbed feedback and a Sennheiser headset microphone. They were required to speak the target phrase 110 times within a certain timeframe visualized on a screen in front of them and were encouraged to keep their speech rate as constant as possible throughout the experiment. Perturbation was applied in phases, whereby the first 20 trials served as a baseline with no perturbation, followed by a ramp phase over 30 trials with increasing perturbation, culminating in a hold phase with 30 trials of maximum perturbation, and followed by a 30 trials after-effect phase with no perturbation again.

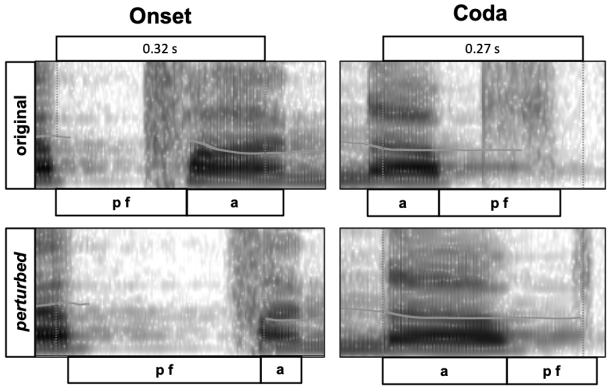
2.3. Temporal real-time perturbation using Audapter

The experiment was conducted in MATLAB using the AUDAPTER software package of Cai et al. [1]. Online real-time perturbation allows to apply perturbations to a predefined substructure in fluent speech, e.g. certain sounds in a syllable. To trigger the perturbation section, AUDAPTER performs an *online status tracking* (OST) based on detection of pre-defined high- and low-frequency weighted intensity thresholds. In this experiment, OST thresholds were predetermined to fit to the word "besser". The end of the OST determines the start of the *perturbation section* (PS) where the manipulation is applied. To estimate the length of the PS (meaning the length of the CCV (/pfa/) and VCC (/apf/) segments), a pretest was performed with every subject in which the mean duration of the produced CCV/VCC sequence was calculated and integrated into the test procedure.

The real-time perturbation needs a short delay of not-noticeable 20ms to process and feed-back the AF to the subject. To maintain this delay over the total duration of the utterance, the temporal manipulation must stretch and compress within the PS by the same amount. Since it is not possible to first compress a sound (because in this case the feedback to be sent back would not have been produced yet), the manipulation will always stretch the first 50% of the PS up to 80%, and compress the second 50% down to 20% of the original length.

Hence in the onset condition ("Pfannkuchen") CC (/pf/) will mostly be stretched and the vowel (/a/) will be compressed, while in the coda condition ("Napfkuchen") the vowel (/a/) will be stretched and CC (/pf/) will mostly be compressed. An example of perturbation for both onset and coda condition can be found in Figure 1.

Figure 1: Example of original signal (baseline, above) and applied maximum perturbation (hold phase, below) in the onset condition (left) and coda condition (right). The time span above marks the perturbation section (0.32s in the onset condition and 0.27s in the coda condition).



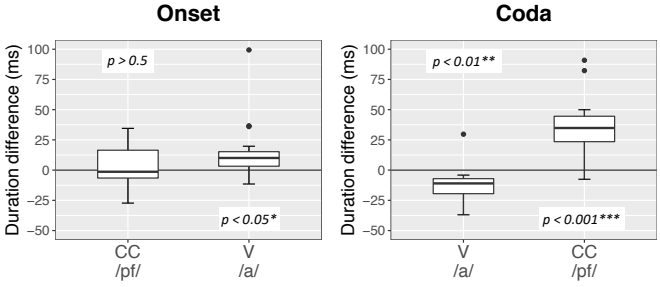
The effect of the temporal AF manipulation was tested with two calculations. Firstly, the produced absolute segment durations in baseline and hold phase were compared to show whether subjects in principle react to a temporal perturbation. Secondly, the expected duration differences in production were set in relation to the amount of perturbation that was applied. This calculation gives insight into the strength of reaction and allows a comparison between onset and coda condition.

3. RESULTS

3.1. Segment durations in baseline vs. hold phase

Linear mixed models were calculated with fixed factor perturbation phase (baseline vs. hold) and subject as random factor over total segment durations of CC (/pf/) and V (/a/) in the onset condition, and V (/a/) and CC (/pf/) in the coda condition. The models indicated significant differences in segment length for the vowel in both conditions, and for CC in the coda condition (see Figure 2).

Figure 2: Differences in produced segment duration between hold phase and baseline (H-B) for onset/coda perturbation (21/19 subjects). Boxes correspond to the first and third quartiles, bars represent the median. Whiskers extend from the hinge to the highest/smallest value no further than 1.5* IQR. Data beyond whiskers are outliers.



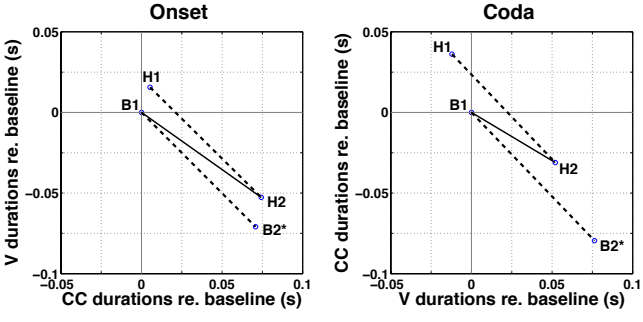
In the onset condition the vowel was produced longer in hold phase relative to baseline (mean difference = 13.8ms, $p < 0.05$, sd = 23.1), and in the coda condition the vowel was produced shorter (mean difference = -11.5ms, $p < 0.01$, sd = 13.4) and CC longer (mean difference = 34. 5ms, $p < 0.001$, sd = 25) in hold phase relative to baseline. Thus, in those three cases subjects adjusted the produced segment duration in the opposite direction to the perturbation. No effect was found for CC in the onset condition (mean difference = 2ms, sd = 17.2), but an effect of testing order was found ($t = 2.85$, $p<0.05$), whereby subjects who performed the onset condition first rather compensated, while the others rather followed the perturbation.

3.2. Compensation relative to perturbation

The analysis of absolute duration differences between baseline and hold phase has shown that subjects compensate for perturbations in the temporal domain in both directions (i.e. compression of the vowel and lengthening of CC in coda condition). To determine how strong the compensation was relative to the applied perturbation, a measure was calculated that takes into account that the perturbation is applied on sounds that may already be produced compensatorily. Further, for inspection of the compensatory behaviour of the perturbed section as a whole (CCV and VCC), a combination of CC and V segments for the onset condition and a combination of V and CC segments for the coda condition were taken into consideration. To ensure a clean comparison between onset and coda condition, only subjects with data in both perturbation conditions were included in following calculations (17 subjects, ø23y, 15 females).

As point of departure we take a normalized two-dimensional coordinate system, wherein the segment durations of the first segment (CC for onset condition and V for coda condition) are on the x-axis and the durations of the second segment (V for onset condition and CC for coda condition) are on the y-axis (for visualization see Figure 3).

Figure 3: Mean durations of both segments per condition normalized to baseline productions (17 subjects). B marks baseline durations, H hold phase durations. B1 and H1 represent the signal spoken by the subject, B2* and H2 the (*simulated) perturbed feedback.



For the following calculations two channels for hold phase (H) and baseline (B) were considered, respectively: the original signal spoken by the subject (1), and the perturbed signal heard by the subject (2). Although there was no perturbation applied in the baseline, a simulation of the signal with perturbation was generated to estimate the maximum perturbation on a signal without compensation (B2*). The durations were normalized to the baseline production (B1), hence B1 is at the zero-crossing for both axes.

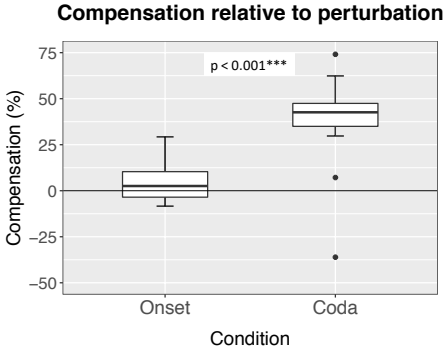
A *mean perturbation* was calculated from the mean of (simulated) maximum perturbation without compensation (Euclidian Distance |B1-B2*|, dashed line) and perturbation on a signal with reaction of the subject (Euclidian distance |H1-H2|, dashed line) (see equation 1). Assuming that subjects intuitively aspire to match the received AF with the intended speech signal through compensation, a closer distance between B1 (spoken and heard signal without perturbation) and H2 (perturbed AF in the hold phase) would mean a stronger compensation. If H2 equals B1 the reaction is interpreted as perfect compensation, meaning that the subject heard the signal he or she intended to speak. The Euclidian distance of |B1-H2| (solid line) was then divided by the *mean perturbation* and scaled to percent values (see equation 2).

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

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

Based on these calculations we observed compensation values between -8% and 29% for the onset condition (mean = 4.9%, sd = 10.5, median = 2.6%), and values between -36% and 74% (mean = 37.9%, sd = 23.7, median = 42.6%) for the coda condition. A negative value results from a following of the perturbation (at least for one of the sounds). A paired t-test was executed to estimate the relation of onset compensation to coda compensation which turned out to be significant, showing a higher compensation in the coda condition (t = 4.48, p = < 0.001, see Figure 4).

Figure 4: Compensatory behaviour for both conditions (17 subjects). Boxplot statistics are the same as in Figure 2.



4. DISCUSSION

In this study the effect of real-time temporal AF perturbation was tested on the timing of segments within specific syllable positions. The preceding analyses showed that subjects do react to manipulations of segment durations and that they mainly compensate in the opposite direction to the perturbation, just as has been found for spatial perturbations. Compensation was bidirectional for the vowel perturbation, resulting in compensatory lengthening and shortening of segments. For the CC segment there was a compensatory lengthening observed in the coda condition, but no compensatory shortening in the onset condition. Hence, the compensation in terms of absolute segment durations was overall stronger for the coda than for the onset condition. An effect of testing order for CC in the onset condition was found, but further consideration of possible effects of testing order must await testing of more subjects.

Typical spatial perturbation studies showed compensation values around 25-30% (for F1 and F2) [5]. In this study we found values of 5% compensation relative to perturbation for onset manipulations and 38% compensation relative to perturbation for coda manipulations. Hence at this stage of our investigations we cannot definitively state whether spatial or temporal perturbation elicits more compensation; however, the observed coda compensation is even higher than averaged spatial compensations.

Significantly stronger compensatory effects were found in the coda condition which gives some support to our assumption that onset clusters are more robust when it comes to a distractor than coda clusters, since the coupling relations in CCV may be more constrained than in VCC [3].

In this investigation we were able to show that subjects react similarly to a perturbation in the temporal domain as they do to perturbations in the spatial domain. To our knowledge compensation to a temporal perturbation of this kind has not been found before. With these results it might be conceivable that temporal properties of speech sounds and syllable structure are stored in a similar manner to acoustic speech targets as described in DIVA; and that AF plays a crucial role for timing of sound and syllable duration.

In the next phase of the investigation, more subjects will be observed and their reactions to a sudden absence of perturbation (after-effect phase) and while perturbation increases (ramp phase) will be analysed.

5. ACKNOWLEDGEMENTS

Work supported by DFG grant HO 3271/6-1. Thanks to Sebastian Böhnke for his help in running the tests.

6. REFERENCES

[1] 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/. *Proc. 8th ISSP* Strasbourg, 65–68.

[2] Cai, S., Ghosh, S. S., Guenther, F. H., Perkell, J. S. 2011. Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *Journal of Neuroscience* 31(45), 16483–16490.

[3] Hoole, P., Pouplier, M. 2015. Interarticulatory Coordination: Speech Sounds. In: Redford, M. A. (eds), *The Handbook of Speech Production*. Wiley: 131–157.

[4] Klein E., Brunner J., Hoole, P. 2018. Which factors can explain individual outcome differences when learning a new articulatory-to-acoustic mapping? In: Fang Q., Dang J., Perrier P., Wei J., Wang L., Yan N. (eds.). *Studies on Speech Production. Proc. 11th ISSP. Lecture Notes in Computer Science* 10733, Cham: Springer 158–172.

[5] MacDonald, E. N., Goldberg, R., Munhall, K. G. 2010. Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am*. 127(2), 1059–1068.

[6] MacDonald, E., Munhall, K. G. 2012. A preliminary study of individual responses to real-time pitch and formant perturbations. *Proc. The Listening Talker* Edinburgh, 32-35.

[7] Morton, J., Marcus, S., Frankish, C. 1976. Perceptual centers (P-centers). *Psychological Review* 83(5), 405–408.

[8] Purcell, D. W., Munhall, K. G. 2006. Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am*. 119(4), 2288–2297.

[9] Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S., S., Houde, J., F. 2016. A New Model of Speech Motor Control based on task dynamics and state feedback. *Proc. Interspeech* San Francisco, 3564–3568.

[10] Saltzman, E., Byrd, D. 2000. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science* 19(4), 499–526.

[11] Tourville, J. A., Guenther, F. H. 2011. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes* 26(7), 952–981.