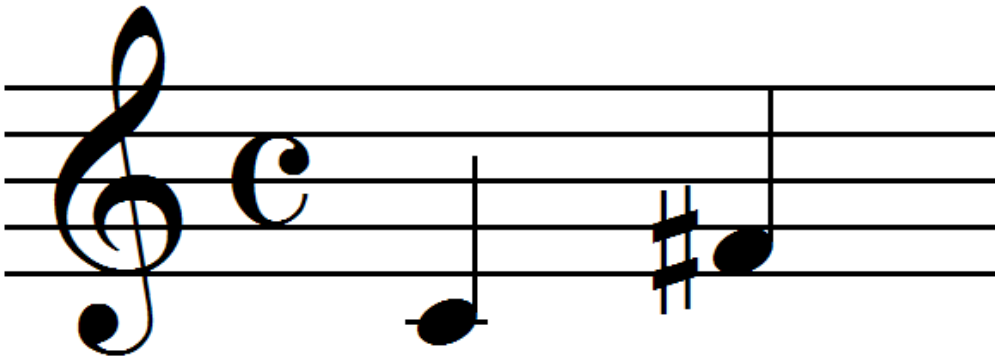
Proceedings of

DiSS 2017

The 8th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
18–19 August 2017

TMH-QPSR
Volume 58(1)



Edited by
Robert Eklund & Ralph Rose

Conference website: http://www.diss2017.org
Proceedings also available at: http://roberteklund.info/conferences/diss2017

Cover design by Robert Eklund
Graphics and photographs by Robert Eklund (except ISCA and KTH logotypes)

Proceedings of DiSS 2017, Disfluency in Spontaneous Speech
Workshop held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 18–19 August 2017
TMH-QPSR volume 58(1)
Editors: Robert Eklund & Ralph Rose
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISSN 1104-5787
ISRN KTH/CSC/TMH–17/01-SE

The time course of self-monitoring within words and utterances

Sieb Nooteboom and Hugo Quené
Utrecht Institute of Linguistics OTS, Utrecht University

Abstract

The within-word and within-utterance time course of internal and external self-monitoring is investigated in a four-word tongue twister experiment eliciting interactional word initial and word medial segmental errors and their repairs. It is found that detection rate for both internal and external self-monitoring decreases from early to late both within words and within utterances. Also, offset-to-repair times are more often of 0 ms in initial than in medial consonants.

Introduction

This paper is about the time course of both prearticulatory and postarticulatory self-monitoring within words and within utterances. We derive and test a few predictions on detecting and repairing segmental errors in different positions in the word and in the utterance. We make the following assumptions on the processes of internal and external detection of segmental errors. These assumptions are taken from the computational model described by Hartsuiker and Kolk (2001) plus a few modifications as suggested by Nooteboom and Quené (2017).

A list of phases in generating segmental speech errors and detecting these errors by self-monitoring may look as follows:

1) Lexical selection.
2) Phonological encoding.
3) Selection of motor plan: 100 ms per lexical item.
4) Execution of motor plan: 100 ms per syllable.
5) Parsing encoded form: 100 ms per lexical form.
6) Comparing error and target form 50 ms.
7) Error detected at least 150 ms after phonological encoding is completed.
8) After error detection in internal speech, both an interruption command is issued and executed (150 ms) and a command to repair is issued.

If a segmental error is not detected in internal speech, it may be detected later in overt speech. In that case parsing and comparing start about 50 ms after overt articulation has started. The time gap between internal and external detection of segmental errors is at least 350 ms according to Hartsuiker and Kolk (2001). Nooteboom and Quené, (2017) found this time gap to be roughly 500 ms.

Ad 1). During speech preparation successive lexical items are activated (cf. Levelt, Roelofs & Meyer, 1999). A lexical item sent to phonological encoding to become a lexical form remains active for a few hundreds of ms. This supports correct and not misspoken lexical forms during further processing.

Ad 2). For each lexical item a prosodic frame, specifying stress pattern and slots for segments, is selected, and segments are selected to fill these slots, leading to a lexical form (Levelt, Roelofs & Meyer, 1999). The buffer for phonological encoding may contain more than one lexical form. At this level segments in similar positions (cf. Nooteboom & Quené, 2015) may interact leading to one or more error forms. The moment when the phonological encoding is completed and the lexical form is forwarded will here be called T1.

Ad 3) and 4). These assumptions on timing are taken from Hartsuiker and Kolk (2001). Together they imply that articulation of a two-syllable word starts some 300 ms after T1.

Ad 5) and 6). These assumptions on timing too stem from Hartsuiker and Kolk (2001). Self-monitoring inspects the encoded forms by parsing each form (100 ms per lexical form) and comparing the parsed form with the still active correct target form (50 ms per form). We assume that error form and correct target form can be simultaneously active and in competition (cf. Nooteboom & Quené, 2017), potentially leading to interactions between correct and error segments in comparable positions and thus to segment replacements or articulatory blends as described by Goldstein et al. (2007) and McMillan and Corley (2010).

Ad 7). Because parsing and comparing together take some 150 ms, an error can be detected not sooner than 150 ms after T1. Let us call the moment of error detection T2. We assume that parsing and comparing in search of a speech error is similar to scanning a word form in a silent phoneme detection task as described by Wheeldon and Levelt (1995). They found that phoneme detection in internal speech takes progressively more time going from early to later phoneme positions. Likewise we assume that whereas segmental errors in initial

position can be detected at 150 ms after T1, detection of segmental errors in later positions takes more time. Assuming that the time needed for scanning a lexical form for speech errors is roughly equivalent to speaking time, T2 will fall substantially later after T1 for later segments than for initial segments, but the time interval between T2 and the moment the error segment is actually spoken will be the same for segments in different positions.

Ad 8). At the moment a segmental error is detected in internal speech, a command to interrupt the process of speaking the form containing the error is issued. According to Hartsuiker and Kolk (2001), execution of an interruption command takes 150 ms. As we have seen above, the process of starting articulation after phonological encoding takes some 300 ms. Now we see that the process of error detection (150 ms) plus speech interruption (150 ms) also takes some 300 ms. Assuming that there is noise in the timing of the various processes involved, it follows that after a segmental error is detected in internal speech, the process of speaking the form containing the error can be interrupted before or after articulation has started. For errors in initial position this implies that the distribution of error-to-offset times is incomplete: Cases in which speech is stopped before articulation has started are invisible. For errors later in the word form, this may lead to utterances for example of the form *ba..bakery* where the internal error may have been *bapery*. Here also the internal error is invisible.

During attempts to repair after internal error detection the correct target form is still active in many cases. If so, repairing can be very fast. After external error detection, however, which happens at least some 350 ms later, the chances are that the correct target has been de-activated. If so, planning a repair will be more time-consuming.

Because in the above view of self-monitoring the average moment of interruption is coupled to the position of the error segment in the word, potentially the proportion of observable errors would remain unaffected by the position in the word. But there is reason to suspect that the observed detection rate is indeed affected by the position in the word. Self-monitoring requires attention (Levelt, Roelofs & Meyer, 1999), and during speaking attention is divided over different processes, for example speech preparation, articulation, but also self-monitoring internal speech and self-monitoring external speech (Hartsuiker, Kolk & Martensen, 2005). In scanning a word form for errors, more and more attention will be needed for preparing and speaking the next word as the end of the current word comes nearer. Therefore one would not be surprised to find that less and less attention is paid to self-monitoring going from earlier to later in the word form. If so, this would lead for example to a predicted difference in detection rate between initial and medial consonant errors:

(1) Detection rate for segmental speech errors will be lower for medial than for initial consonant errors.

A similar argument can be made with respect to position of the word in the utterance. From the first word on, the number of possible interactions to be detected increases, decreasing the amount of attention for detecting later speech errors. From this we predict that:

(2) Detection rate for segmental speech errors will decrease with position of the word in the utterance from early to late.

According to Hartsuiker and Kolk (2001) both the interruption and the command to produce a repair are triggered by the error detection and executed in parallel. Therefore a difference between initial and medial consonants in the timing of a repair as measured in offset-to-repair times has already been compensated by the later moment of interruption. For this reason we predict that there is no difference between earlier and later segmental errors in offset-to-repair times:

(3) Offset-to-repair times are equal for initial and medial consonant errors.

A similar argument can be made for segmental errors in earlier and later words in the utterance:

(4) Offset-to-repair times do not depend on the position of the words in the utterance.

Experiment

We have elicited segmental speech errors in tongue twisters, each tongue twister consisting of 4 two-syllable CVCVC words (Shattuck-Hufnagel, 1992). There were 4 conditions, one with only strong-weak (Sw) stress patterns, one with only weak-strong (wS) stress patterns, one with the sequence Sw wS wS Sw, and one with the sequence wS Sw Sw wS. In each condition there were 48 tongue twisters, 24 meant to elicit interaction between initial or medial consonants by consonant repetition (as in *kennis gekkie gele kater* for initial position), and 24 without consonant repetition for that position. Initial and medial positions were controlled for the opportunities for interaction (Nooteboom & Quené, 2015). Thirty participants were asked to speak each tongue twister 6 times, 3 times reading from a screen and 3 times from memory (Shattuck-Hufnagel, 1992). All speech was transcribed and coded as to segmental speech errors, keeping the four word positions and within-word segmental

positions separate. For all repaired single segment word initial and word medial speech errors onset-to-cutoff intervals (from word onset to interruption), error-to-cutoff intervals (from onset of error segment to interruption), and offset-to-repair intervals (from interruption to repair onset) were measured.

We focus on consonant errors in initial and medial positions, because other positions were not controlled for the numbers of opportunity for interaction. We exclude all cases in which specific errors, due to hysteresis, were repetitions of the same error by the same speaker. In the current experiment we wished to elicit segmental errors in different positions in the word and in different words in the utterance. We were predictably punished by much variation in error-to-cutoff times, leading to considerable overlap between internally and externally detected repaired errors. Because of this, there is no possibility to estimate the form of the underlying distributions of internally and externally detected errors, and thus no possibility to separate between internal and external error detection, as was done in Nooteboom and Quené (2017).

The detection rates of speech errors are summarised in Figure 1. One may note that the total number of single segment errors is lower for medial than for initial position.

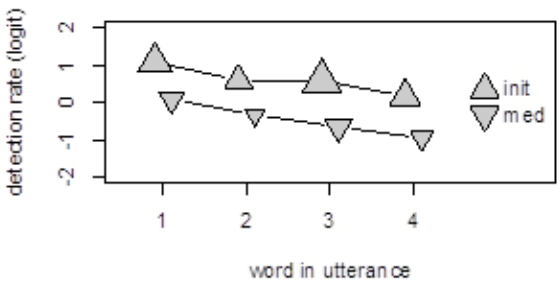


Figure 1: Detection rates of single-segment errors, broken down by word position and by within-word position of the error. Symbol size corresponds to numbers of errors in each cell.

Binomial detection status of each valid error was analysed by means of a GLMM (Quené & Van den Bergh, 2008), with position (initial vs medial) and word number (one to four) as two fixed factors, with response number (1,4,5,6 vs. 2,3) as an additional fixed factor, and with participants and items as random intercepts. Position and word number were also added as random slopes over participants. Interactions were dropped from the GLMM as these did not improve the model fit, according to Likelihood Ratio tests.

The GLMM shows a significant main effect of consonant position, with a significantly lower detection rate for medial than for initial consonant errors (beta=-1.169, Z=-7.1, p<.0001). We also found a significant main effect of word position (beta=−0.320, Z=−4.1, p<.0001): Detection rate decreases within utterances from earlier to later words. We explain the decrease of detection rate from initial to medial consonant error positions, and from earlier to later words in utterances as reflecting differences in attention available for self-monitoring.

We have also measured onset-to-offset times (from word onset to interruption) and error-to-offset times (from onset of error segment to interruption) for all initial and medial segmental errors. Of course, for initial segments onset-to-offset times and error-to-offset times are identical. For medial consonants the moment of interruption on average falls 155 ms later than for initial consonants. However, with Hartsuiker and Kolk (2001) we have assumed that both the command to interrupt speech and the command to plan a repair are triggered by the moment of error detection and executed in parallel. If the interruption after detecting a medial consonant error is 155 ms later than the interruption after detecting an initial consonant error, then also the initiation of generating the repair should be 155 ms later for medial than for initial consonant errors. As the offset-to-repair times were measured from the moments of interruption, this difference of 155 ms is already taken into account. Therefore a priori we expect for internally detected errors no difference in offset-to-repair times between initial and medial positions. Also for externally detected errors there is no reason to expect a difference in offset-to-repair times. The distributions of offset-to-repair times are shown in Figure 2.

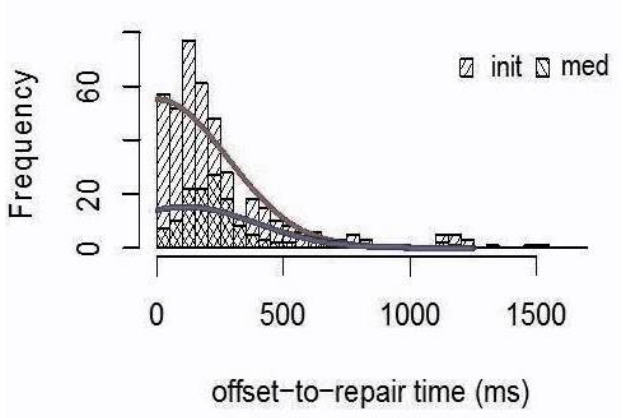


Figure 2: Histograms and fitted truncated gaussian distributions of offset-to-repair times, for initial and medial consonants (see text).

Figure 2 shows that offset-to-repair times are truncated at 0 ms, for errors involving initial (48/426 errors) and medial consonants (5/140 errors). The untransformed error-to-repair times

were therefore analyzed using regression techniques for such truncated distributions (Croissant & Zeileis, 2016). Error times exceeding 1000 ms (3.7% of total) were ignored for this analysis. Since models including word position as a predictor could not be estimated reliably, we focus here on consonant position as the only predictor. Random intercepts and slopes of participants and of item sets were also ignored, as these could not be estimated.

The optimal model for the truncated error-to-repair times showed a non-significant effect of consonant position, according to a Likelihood Ratio test (χ^2=2.72, df=1, p=.099) of the optimal model compared to a null model. For initial consonants, the estimated mean of error-to-repair times is 0 ms [with bootstrapped 95% confidence interval (−210, +65) over 1000 iterations], whereas for medial consonants, the estimated mean error-to-repair time is +113 ms [with 95% confidence interval of (−78,+181)]. It seems possible that there is a difference between initial and medial consonants, but that it cannot be demonstrated due to a lack of statistical power. This is confirmed by a separate analysis of the numbers of immediate (offset-to-repair time of 0 ms) vs non-immediate repair (offset-to-repair time > 0 ms). These were analyzed by means of a GLMM with the same fixed and random predictors as before. The odds of an immediate repair were significantly lower for medial consonants than for initial consonants (beta=-1.237, Z=−2.424, p=.0154), and the odds were significantly higher for words in the third position than for words in the baseline first position (beta=+0.918, Z=2.251, p=.0244). The interaction between consonant position and word position was not significant (p=.326, Likelihood Ratio Test).

The cases in which the offset-to-repair time is 0 ms demonstrate that a repair is ready to be spoken at the moment of interruption. Our results show that this happens significantly more often after word initial consonant errors than after medial consonant errors. This can hardly be explained from the small and insignificant difference in error-to-offset times between initial (217 ms) and medial (198 ms) consonant errors: The time for preparing a repair after error detection and before speech is interrupted is roughly similar for the two consonant positions. Therefore our results suggest that preparing a repair takes more time for medial than for initial consonant errors. Earlier, we have suggested that the amount of attention available for self-monitoring is less for medial than for initial consonants, causing a lower error detection rate for medial than for initial consonant errors. We now suggest that less attention not only lowers detection rate, but also leads to slower error repair.

Conclusion

Main findings are (1) Detection rate of errors is lower in medial than in initial consonants; (2) Detection rate decreases from the first to the last word in four-word utterances; (3) Relatively, there are many more offset-to-repair times of 0 ms for initial than for medial consonants.

Findings 1) and (2) suggest that attention for self-monitoring decreases from early to late both within words and within utterances.

Finding (3) shows that a repair is often available before speech is interrupted, and that repairing medial errors takes more time than repairing initial errors. These results confirm the computational model of Hartsuiker and Kolk (2001) and shed further some light on the time course of self-monitoring.

References

Croissant, Y. & A. Zeileis. 2016. *truncreg: Truncated Gaussian Regression Models*. R package version 0.2-4. Available at: https://CRAN.R-project.org/package=truncreg.

Goldstein, L., M. Pouplier, L. Chen, E. Saltzman & D. Byrd. 2007. Dynamic action units slip in speech production errors. *Cognition* 103:386–412.

Hartsuiker, R. J. & H. H. J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology* 42:113–157.

Hartsuiker, R. J., H. H. J. Kolk & H. Martensen. 2005. Division of labor between internal and external speech monitoring. In R. Hartsuiker, Y. Bastiaanse, A. Postma & F. Wijnen (eds.), *Phonological encoding and monitoring in normal and pathological speech*. Hove: Psychology Press, 187–205.

Levelt, W. J. M., A. Roelofs & A. S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22, 1–75.

McMillan, C. T. & M. Corley. 2010. Cascading influences on the production of speech: Evidence from articulation. *Cognition* 117:243–260.

Nooteboom, S.G. & H. Quené- 2015. Word onsets and speech errors. Explaining relative frequencies of segmental substitutions. *Journal of Memory and Language* 78:33–46.

Nooteboom, S.G. & H. Quené. 2017. Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback, *Journal of Memory and Language* 95:19–35.

Quené, H. & H. Van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. Journal of Memory and Language 59:413–425.Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering. *Cognition* 42:213–259.

Wheeldon, L. R. & W. J. M. Levelt. 1995 Monitoring the time course of phonological encoding. Journal of *Memory and Language* 34:311–334.