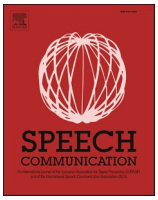
Speech Communication 158 (2024) 103043

Contents lists available at ScienceDirect

Speech Communication

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



Some properties of mental speech preparation as revealed by self-monitoring

Hugo Quené *, Sieb G. Nooteboom

Institute for Language Sciences, Utrecht University, Trans 10, NL-3512 JK Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Self-monitoring
Speech preparation
Speech errors
Auditory perceptual targets
Motor commands

ABSTRACT

The main goal of this paper is to improve our insight in the mental preparation of speech, based on speakers' self-monitoring behavior. To this end we re-analyze the aggregated responses from earlier published experiments eliciting speech sound errors. The re-analyses confirm or show that (1) "early" and "late" detections of elicited speech sound errors can be distinguished, with a time delay in the order of 500 ms; (2) a main cause for some errors to be detected "early", others "late" and others again not at all is the size of the phonetic contrast between the error and the target speech sound; (3) repairs of speech sound errors stem from competing (and sometimes active) word candidates. These findings lead to some speculative conclusions regarding the mental preparation of speech. First, there are two successive stages of mental preparation, an "early" and a "late" stage. Second, at the "early" stage of speech preparation, speech sounds are represented as targets in auditory perceptual space, at the "late" stage as coordinated motor commands necessary for articulation. Third, repairs of speech sound errors stem from response candidates competing for the same slot with the error form, and some activation often is sustained until after articulation.

1. Introduction

This paper focuses on the general question how speech sounds are organized and represented in internal speech. Although mental preparation of speech is inherently difficult to observe, speakers' self-monitoring for and detection of experimentally elicited errors, with control over the elicited speech errors, may offer valuable insights. Speech sound errors (experimentally elicited or spontaneous errors) can be detected by the speaker during at least two distinct stages of speech production (Levelt, 1983, 1989; Levelt et al., 1999; Hartsuiker and Kolk, 2001) which are further discussed below. At the first of these stages, an error is detected "early", that is, in the speaker's internal speech, which equals the speaker's internal plan of speech sounds to be articulated. At the second stage, an error is detected "late", in the speaker's overt speech (hearing one's own error) or in somatosensory feedback from the articulators (sensing one's own error).

In this paper we first confirm and expand earlier results on self-monitoring for speech sound errors, experimentally answering three different questions outlined below. More importantly, we then attempt to derive properties of successive stages in the mental preparation of speech from our insights in self-monitoring. Of necessity this cannot be done without some speculation.

Our first experimental question concerns the difference in detection timing of elicited speech errors that are detected "early" vs. "late" (cf. Nooteboom and Quené, 2008, 2017) using the pooled data reported in the latter two articles. Secondly, knowing that fewer errors are typically observed if the phonetic contrast between the intended sound and the elicited error is relatively large (Dell, 1986), we develop the additional hypothesis that a larger phonetic contrast also yields relatively *more detections,* especially "early" detections, and our second aim is to investigate this new hypothesis. The third aim of this paper is to re-investigate several predictions (from Nooteboom and Quené, 2020) about the activation levels of repair candidates using the pooled data sets.

The re-analyses reported in this paper are relevant for two reasons. Data on self-monitoring of elicited speech sound errors are scarce: relatively few of such elicitation experiments have been reported (to be discussed in the next section below), and the numbers of elicited errors and of detections in those experiments are appallingly low. Pooling data from multiple comparable experiments, and using advanced statistical methods, will increase statistical power and robustness of the results. Moreover, joint investigation using the same pooled data set to answer

* Corresponding author.
E-mail address: h.quene@uu.nl (H. Quené).

https://doi.org/10.1016/j.specom.2024.103043
Received 23 May 2023; Received in revised form 31 October 2023; Accepted 8 February 2024
Available online 9 February 2024

related questions regarding (1) patterns in detection timings, (2) effects of phonetic contrast among target and competitor, and (3) activation levels of repair candidates, will allow for more coherent interpretations and conclusions about the mental preparation of speech.

After a speech error is made in internal speech, the error form containing the speech error is lined up for being transduced into a sequence of coordinated motor commands for articulation. Meanwhile, self-monitoring of *internal* speech detects the error, but internal error detection and interrupting the flow of speech for repairing takes at least the same amount of time as triggering and executing the motor commands leading to articulation of the error form (cf. Hartsuiker and Kolk, 2001). This implies that the earliest moment where speech can be interrupted after internal detection of a speech sound error coincides with the initiation of speaking the error form. In many cases initiation of speaking the error form comes much later than interruption, for example because error detection is slow or because interruption is delayed due to the absence of an available repair. This, after internal error detection, leads to a Gaussian distribution (in log ms) of error-to-interruption times between 0 ms and many hundreds of ms. Alternatively, an error may be detected only later, that is, after somatosensory targets for articulation have been activated or articulation has started. This leads to a Gaussian distribution of error-to-interruption times between c. 300 ms and more than 1000 ms (Nooteboom and Quené, 2017). This scenario explains the bimodality of the distribution of error-to-interruption times.

In itself, the existence of two stages of self-monitoring for speech errors provides only little information on the organization of speech preparation. With respect to different stages of the preparation of segmental speech, it is conceivable that there is only a single level of internal speech, and that the second stage of self-monitoring is directed at the speaker's own overt speech. Such a single level of segmental internal speech seems to be assumed for example by articulatory phonology (Browman and Goldstein, 1992). However, some theories of speech preparation assume at least two distinct stages of preparation of speech sounds (cf. Levelt et al., 1999; Guenther, 2016). In that case, "late" self-monitoring could still be directed only at overt speech, as apparently believed by Levelt et al. (1999; see also Hartsuiker and Kolk, 2001; Nozari et al., 2011), but it might in principle also be directed at the "late" stage of speech preparation, where coordinated motor commands are planned for controlling articulation. Finally, self-monitoring could be directed at articulation itself, employing somatosensory feedback from the articulators (cf. Hickok, 2012; Lackner and Tuller, 1979; Nooteboom and Quené, 2017; Pickering and Garrod, 2013).

In pursuing our quest for properties of the mental preparation of speech further, we draw on 6 data sets obtained in experiments eliciting segmental speech errors (SLIP: Spoonerisms of Laboratory Induced Predisposition, cf. Baars and Motley, 1974), reported in the last 15 years. We focus in this paper on some questions that, even after many decades of research by many researchers, still seem to be controversial, and that may be addressed by re-analyzing the aggregated data from these six data sets. (In the past two decades we have run several SLIP experiments, yielding more responses than those included in the six data sets re-analyzed below; however, the differences between those experiments, in experimental details and materials, would make it difficult to interpret the results from such an enlarged data set in a coherent way.)

2. Theoretical background

Influential theories of the mental processes involved in speech production have been proposed by, among others, Dell (1986), Goldstein et al. (2007), Guenther (2016); Hickok (2012), Levelt et al. (1999) and Pickering and Garrod (2013). Theories particularly focusing on self-monitoring for speech errors during speech production have been proposed by, among others, Gauvin and Hartsuiker (2020), Hartsuiker and Kolk (2001) and Nozari et al. (2011). Here we will briefly mention some important differences between these theoretical accounts, explain what in each case our own standpoint is, and which questions we will attempt to answer in this paper. We notice that some of the old controversies still exist, as for example shown in a relatively recent exchange of ideas on comprehension-based monitoring versus production-based monitoring (Roelofs, 2020a; Nozari, 2020; Roelofs, 2020b).

A major difference between Dell (1986) and Levelt et al. (1999) is that Dell proposes that there is *immediate feedback* between levels of speech production, notably between the lexical level and the segmental level, whereas Levelt et al., proposing serial processing in speech production, explicitly claim that there is no such immediate feedback (as opposed to feedback through listening to one's own speech) between successive levels of speech production. Since then, some evidence has shown that feedback between lexical selection and processing of speech sounds in speech production is real (e.g. Hartsuiker et al., 2005; Nooteboom and Quené, 2008). Both of these latter papers investigated lexical bias in speech sound errors, i.e. the tendency that, other things being equal, speech sound errors create real words instead of nonwords, and both papers suggest that lexical bias is caused by both immediate feedback and self-monitoring.

In contrast with Dell (1986), the model by Levelt et al. (1999) also prohibits regular *cascading* of information from the lexical level to speech sound processing. For example, in a buffer memory containing a sequence of speech sounds waiting to be articulated, for each slot in the sequence of sounds only a single unit or segment is passed on from the previous level. This implies that there can be no conflict between simultaneously activated segments, passed on from the previous level, and competing for the same slot. However, such conflict/competition was reported by Goldstein et al. (2007) and McMillan and Corley (2010) to occur frequently in experiments eliciting articulatory blending between interacting initial consonants in pairs of monosyllabic words. An example is the pair *cop top*, where during articulation of the initial [k] a partial movement of the tongue tip can be observed; this gesture indicates that planned articulation of the interfering [t] is also active. In our own experiments eliciting segmental speech errors in Dutch, we too have observed that cases in which initial consonants rapidly alternate, as in *feit goud* > *gfgeitfout, tand veeg* > *tftantfeeg, bijl geit* > *g*[ɛ]*bgbijlgeit, duit vast* > *dvduitvast, paf kies* > *p*[ə]*k*[ə]*pfafkies,* occur too frequently to be ignored (cf. Nooteboom and Quené, 2017). Such cases also suggest that conflicting speech sounds or segments can be simultaneously active, competing for the same segmental slot. We here assume that both immediate feedback and cascading of information are normal features of speech production. It may be noted that the rapid alternation of competing speech sounds in the examples above is virtually limited to the initial consonants of word-like forms. This suggests that conflict between segments competing for the same slot is constrained by competition between word-like forms (called "phonological words" by Levelt et al., 1999).

As mentioned in the Introduction above, most current theoretical accounts of speech production acknowledge that speech errors can be detected by the speaker at least at two stages (Levelt, 1983, 1989; Levelt et al., 1999; Hartsuiker and Kolk, 2001): "early" in internal speech or "late" in overt speech or in articulation. Internal speech is, according to current terminology, a stage of speech preparation, not to be confused with "inner speech" which would be silent, cf. Oppenheim and Dell, 2008. Of course, in a sense both internal speech and inner speech are silent, because internal speech is not yet articulated. The main difference is that internal speech is being prepared for articulation whereas inner speech is to remain silent. The work by Oppenheim and Dell (2008) suggests that certain low level phonetic features, present in internal speech, are absent in inner speech. That speech errors can be detected both "early" and "late" was for example proposed by Levelt (1983; 1989; see also Levelt et al., 1999). Levelt observed that fragments of spoken error forms such as *v* in *v. horizontal* can be so short that it seems highly unlikely that interruption was triggered by the speaker's detection of the overt error fragment. Blackmer and Mitton (1991) observed that very brief error fragments, after interruption, are often followed by very short

interruption-to-repair times that may even be 0 ms. An example would be *bdarn bore* (here the *b* of *bore* supposedly was erroneously anticipated and therefore came to interact with the *d* of *darn*, leading to an audible anticipation of the *b* of *bore,* apparently replacing the *d* of *darn* for a brief moment). As observed by Blackmer and Mitton (1991), in such cases both interruption and repair must have been planned before speech is initiated.

Levelt et al. (1999) proposed that before articulation, the contents of a buffer memory that we might equate with "internal speech", are passed on to the comprehension system. They assumed that "early" and "late" detection would both take place in the speech comprehension system: Self-monitoring of internal speech would be directed at the output of speech preparation, forming a buffer memory containing a sequence of speech segments to be articulated. The contents of this buffer memory form the input for both the comprehension system, leading to self-monitoring internal speech, and for articulation, leading to the acoustic wave form of speech, that via audition also is analyzed by the speech-comprehension system. Self-monitoring of both internal and external speech would be "comprehension-based". Together these two pathways form the so-called "dual perceptual loop" of self-monitoring (cf. Hartsuiker and Kolk, 2001). "early" and "late" detection seems to refer to detection of errors at different, successive stages of the mental preparation of speech. If so, this would allow us to learn something from the relative timing of "early" and "late" error detection about the relative timing of these successive stages of speech preparation. In this paper we make this assumption. However, from the kind of data we work with it is not clear that the difference between two stages of error detection measured in ms means that this difference reflects a difference in timing between stages of the mental preparation of speech that can also be expressed in ms. Possibly, future research employing brain imaging techniques can reveal more about the timing of different stages of speech preparation. We also wish to point out that, in focusing on speech sound errors only, we are not in a position to draw conclusions on higher order (lexical, syntactic, semantic, pragmatic) errors.

In being comprehension-based, the theory by Levelt et al. (1999) deviates from earlier proposals that assume that speech errors can be detected during grammatical and phonological processing preparing speech production (cf. Laver, 1973; MacKay, 1987). Such models are often called "production-based" (cf. Postma, 2000). A different proposal of production-based self-monitoring was made by Nozari et al. (2011; see also Nozari and Pinet, 2020). These authors took a cue from neuro-cognitive research on perceptual conflict and cognitive control reported by a.o. Botvinick et al. (2001) and Yeung et al. (2004). Nozari et al. (2011) propose that conflict between lexical units or speech sounds that during speech preparation compete for the same slot in the sequence being prepared, triggers a cognitive control center. This cognitive control center, in turn, may trigger action that leads to preventing or repairing a speech error. An essential aspect of Nozari's account is that units competing for the same slot remain active during successive levels of speech preparation. We agree with that. We have also argued that competition between candidate word forms is often sustained even after a speech error has become overt, then becoming a major source of repairs (Nooteboom and Quené, 2020). Gauvin and Hartsuiker (2020) go further in the conflict-based direction of Nozari et al. (2011). They propose a computational model for the conflict-based detection and correction of semantic errors. Higher conflict leads to a higher neurotransmitter-derived temporary boost of processing, which may lead to interruption and repair. Gauvin and Hartsuiker (2020) do not discuss speech sound errors.

With respect to Guenther (2016), we point out that most of his theory, implemented in DIVA (Tourville and Guenther, 2011), concerns the neurological processes involved in the production of single speech segments (either single speech sounds or combinations of more than a single speech sound). In DIVA there is no level of representation where speech sounds in different positions, for instance initial consonants of different words, can interact with each other. DIVA does involve a speech sound map, containing targets in auditory perceptual space. Auditory targets are transduced into somatosensory (tactile & proprioceptive) targets that trigger in a feedforward way the coordinated motor commands necessary for articulation. Feedback control is most important for learning the appropriate somatosensory targets. In fluent adult speech, feedback control would be too slow to correct erroneous articulatory movements. Generally, in fluent adult speech the production of speech sounds is, according to Guenther c.s, mainly driven by feedforward control. In GODIVA (cf. Bohland et al., 2010; Guenther, 2016) they do give an account of the internal representation of sequences of segments, but do not consider the commitment and detection of discrete speech errors by self-monitoring. Yet the relevance of their work for our current enterprise is that Guenther c.s. explicitly and convincingly argue that speech sounds are represented as targets in auditory perceptual space before they are transformed into motor commands for articulation.

Before we turn to our research questions below, we wish to point out that this study is limited to the role of self-monitoring in the detection and repairing of segmental speech sound errors. We have no data on the role of self-monitoring for other purposes, such as the control of speech quality, pitch and loudness, and the prevention of speech errors before they are even committed internally. However, the existence of speech errors that are interrupted after very brief fragments of speech, close to zero ms, suggests that some speech errors that are committed internally are suppressed before becoming overt.

3. Research questions

The above observations lead us to three questions about speech preparation and self-monitoring. The *first* question concerns the difference in detection timing of speech sound errors detected "early" and "late". What is the time delay between these two stages of self-monitoring? Nooteboom and Quené (2017) have reported a bimodal distribution of error-to-interruption intervals, allowing tentative separation of the two stages of detection, with estimated average time delays of 498 ms (2017, Expt. 1) and 474 ms (2017, Expt. 2) respectively. Here, including error-to-interruption intervals of two similar experiments eliciting speech sound errors (viz. the experiments reported by Nooteboom and Quené, 2008), we may obtain a better and more robust estimate of the difference in detection than in the two separate experiments of Nooteboom and Quené (2017). This first question is relevant (a) because the resulting classification of detected errors as being detected early or late (or not detected at all) will be more robust, and (b) for the following theoretical reasons not considered by Nooteboom and Quené (2008; 2017). If speech sounds would be represented in internal speech in terms of articulatory gestures, as proposed e.g. by proponents of articulatory phonology (Browman and Goldstein, 1992; Goldstein et al., 2007) then one would predict a relatively fast transduction from internal speech to articulation, with only a relatively short time delay. Moreover, the difference between the fastest "early" detection in internal speech and the fastest "late" detection in overt speech or in articulation would also be short. But if, by contrast, speech sounds would be represented in internal speech as targets in auditory perceptual space, as proposed by e.g. Guenther (2016), then one would predict a relatively slow transduction from internal speech to articulatory gestures, with a relatively long delay.

As will be clear later, we believe that "early" detection and "late" detection differ in the stage of speech preparation at which the errors are detected, corresponding for example to the "phonological" and "phonetic" processing in the theory proposed by Levelt (1989) and Levelt et al. (1999). The terms "phonological" and "phonetic" are rather abstract. We ask whether "phonological" is closer to auditory perceptual properties of speech sounds and "phonetic" closer to articulatory properties of speech sounds.

If we distinguish between elicited speech sound errors detected "early" or "late" (note that only about half of the experimentally elicited

speech sound errors are detected at all; cf. Nooteboom, 2005; Nooteboom and Quené, 2008), then the second question arises: *why* are some speech errors detected "early", others "late", and others again not at all? We hypothesize that the odds and speed of detection are affected by the *strength of phonetic contrast* of the speech sounds interfering as target and competitor. Errors involving a stronger phonetic contrast, e.g. between [p] and [s], occur less frequently than errors involving a weaker contrast, e.g. between [p] and [t], or between [p] and [b] (Dell, 1986; Nooteboom, 1969; Shattuck-Hufnagel and Klatt, 1979). In our view, this is due to the generally smaller interference (*weaker* competition) between target and competitor in strong-contrast errors, as compared to weak-contrast errors. The predicted consequences for *detection* of the error vary with the two theoretical accounts of self-monitoring outlined above. According to a perception-based view of self-monitoring, the higher salience of strong-contrast errors should facilitate (early) detection of such rare errors. According to a production-based view, however, the lack of competition between target and competitor should not facilitate but impede (early) detection. Here, we test the predicted interaction between strength of phonetic contrast, and detection category, using data from multiple experiments and using advanced statistical methods.

Alternatively, it cannot be excluded that "late" detected errors were also committed at a later stage of speech preparation. For example, errors can occur in internal speech where speech sounds are coded as targets in auditory perceptual space, but in principle errors could also be committed on the level of somatosensory targets necessary to activate motor commands for articulation (cf. Guenther 2016). However this may be, we predict that weak-contrast errors are more often than strong-contrast errors detected "late". Likewise, we predict that weak-contrast errors are more often than strong-contrast errors not detected at all. (The reader may observe that if indeed (a) there are two successive stages of error detection, and (b) the speech sounds are coded as auditory perceptual targets at the "early" stage and as "articulatory" targets at the "late" stage, then one would expect some differences in the interactions between speech sounds at these two stages; this issue will not be pursued here because the current data do not allow the relevant analysis).

Thirdly, we return to the questions investigated by Nooteboom and Quené (2020, their Q3 and Q4) about the repair of the elicited speech sound error. We hypothesize, with others (see e.g. Nozari and Pinet, 2020, and, for semantic errors, Gauvin and Hartsuiker, 2020), that the intended, non-erroneous response is still available in internal speech as a repair candidate, especially after elicited errors detected early. Also, the model by Gauvin and Hartsuiker on repairs of semantic errors assumes that repairs may stem from boosting the correct candidate competing with the candidate error form. From this, we predict (a) that the most probable repair is the correct (target) response. Presumably, activation of the target and competitor decreases during the delay between "early" and "late" error detection (especially if the time delay would be relatively long); we therefore also predict (b) that repairs after "early" detection are more often correct than those after "late" detection, and (c) that the *interruption-to-repair* times too are shorter after "early" than after "late" detection. Re-investigation of these predictions is relevant because the pooled data sets and more advanced statistical methods decrease the risk of Type II errors, which may have been committed by Nooteboom and Quené (2020).

In sum, we will attempt to answer three questions about self-monitoring by re-analyzing our corpus of responses in six data sets from experiments in which speech sound errors and their repairs were elicited. The questions are the following:

1. What is the delay between "early" and "late" detection of elicited errors?
2. Why are some speech errors detected "early", others "late" and others again not at all?
3. Where do repairs of segmental speech errors come from?

We will also discuss the resulting properties of self-monitoring for speech errors in relation to more general properties of the mental preparation of speech. But first we will describe the corpus used to find answers to our three questions on self-monitoring.

4. Corpus

We attempt to answer the above questions on the basis of responses in experiments using the so-called SLIP technique (Baars and Motley, 1974). The SLIP technique works as follows: Participants are successively presented visually, for example on a computer screen, with priming precursor word pairs such as *dove ball, deer back, dark bone*, followed by a target word pair *barn door,* all precursor word pairs to be read silently. On a prompt, for example a buzz sound or a series of question marks ("??????''), the last word pair seen, i.e. the target word pair, in this example *barn door*, has to be spoken aloud as soon as possible. Intervals between precursors and between the last precursor and the target stimulus word pair are in the order of 1000 ms, as is the interval between the test word pair and the prompt to speak. Due to the priming of an exchange between initial consonants by the precursor word pairs, every now and then the participant will mispronounce a word pair like *barn door* as *darn bore*.

For us, an advantage of this technique to elicit segmental speech errors is that the timing of responses can be rather strictly controlled. A disadvantage is that the method is not very effective in eliciting the desired speech errors. To more or less overcome the limitation created by this disadvantage we have selected from our SLIP experiments done since 2005 (described in Nooteboom and Quené, 2008, 2017) six different experimental sets of data that are in many respects comparable. Nooteboom and Quené (2017) described two SLIP experiments, with each experiment having two conditions, viz. with or without computer-generated so-called "pink noise" of 87 dB SPL(A), intended to disable auditory feedback. For the current purpose we have kept these conditions with and without masking noise separate as different data sets. Data sets from earlier experiments were not included because differences between experiments, for example with respect to the kind of stimuli used and the task of the speakers, were too great to make these data sets comparable with the current ones.

Of course, there were also further differences related to the specific purpose of each experiment. We will now summarize the properties of these experiments that were the same and those that were different. In all 6 experiments we employed the SLIP technique to elicit specific exchanges between initial consonants of two CVC words in Dutch. Stimulus presentation in all experiments was according to the scheme in Table 1, with word pairs for both precursors and test stimuli, consisting of the two CVC words, being presented visually in the centre of a computer screen.

Each speaker was tested individually in a sound-treated booth. The timing of visual presentation on a computer screen was computer controlled. The initial consonants of priming word pairs and target word pairs were chosen from the set /f, s, χ, v, z, b, d, p, t, k/; the contrasts among these initial consonants in the word pairs will be described

Table 1
Example of a test stimulus item together with its precursor word pairs, the prompt for speaking the last word pair seen (see procedure) and the targeted spoonerism. All CVC words are Dutch.

precursor 1	*bouw jool*
precursor 2	*lijf deed*
precursor 3	*koet pop*
precursor 4	*kuur poet*
precursor 5	*kas piet*
test stimulus	*paf kiep*
prompt	*??????*
elicited spoonerism	*kaf piep*

Table 2
Some differences between the 6 data sets.

Data set	Experiment	Nr of participants (sample)	Nr of baseline stimuli	Nr of test stimuli	Nr of filler stimuli
Data set 1	2008 E1	102 (A)	72	72	46
Data set 2	2008 E2	102 (B)	72	72	0
Data set 3	2017 E1	106 (C)	0	55	23
Data set 4	2017 E1	106 (C)	0	55	23
Data set 5	2017 E2	123 (D)	0	110	46
Data set 6	2017 E2	123 (D)	0	110	46
Sum		433	144	474	184

below. In each experiment, the first trial consisting of 5 precursors plus one test (or filler) stimulus word pair was preceded by seven such trial sets that served as warm-up. These were discounted in the analysis. Precursor word pairs and target word pairs (filler, baseline or test; the latter were to be spoken) were presented consecutively, each word pair being presented for 900 ms with blank intervals of 100 ms in between. After the final word pair of each trial a "??????'' prompt, meant to elicit pronunciation of the last word pair seen (the test, baseline or filler stimulus containing the target word pair), was visible during 900 ms and was then immediately followed by a blank screen combined with a loud buzz sound, both of 100 ms duration. Speakers were encouraged to speak the target stimulus before this buzz sound started. In five of the six experiments the blank screen following the "??????'' prompt was immediately followed by a cue consisting of the Dutch word for *repair*, visible during 900 ms and again followed by a blank screen with 100 ms duration. Only in data set 2 this cue for eliciting a repair was left out.

Test stimuli were always preceded by 5 precursors, the last three of which primed an exchange of initial consonants. Filler stimuli were not primed for exchanges and were preceded by 0, 1, 2, 3 or 4 precursors. This made it impossible for participants to anticipate the test stimulus by counting precursors. Speakers were instructed to pronounce the last word pair seen before the "??????" prompt as soon as possible, and before the loud buzz sound. This put time pressure on the speakers. Responses to test and baseline stimuli were transcribed by listening with visual feedback and were also analyzed acoustically, mainly for durational measurements, using different versions of Praat (cf. Boersma and Weenink, 2016). Responses to filler stimuli were neither transcribed nor analyzed.

The main differences between experiments are tabulated in Table 2. There we present number of speakers, baseline stimuli, test stimuli and filler stimuli per experiment.

Since data sets 3 and 4 refer to different sections of a single SLIP experiment, the sampled participants for these data sets are identical (as indicated in Table 2); the same applies to data sets 5 and 6. The resulting total number of 433 participating individuals could even be an overestimation of the sample size, because some of these individuals have in fact participated in multiple of our SLIP experiments. (We cannot determine this overlap in participants across experiments because the responses have been anonymized for privacy reasons, but we estimate it to affect about 10 % of the participating individuals. In the analyses reported below, we will ignore this overlap of participants across experiments.) The absence of filler stimuli in Experiment 2 implied that the target stimulus (to be spoken aloud as soon as possible) always was preceded by 5 precursors. In these cases the priming of an exchange was by all 5 precursors instead of the last 3.

The first two data sets also included a baseline condition: the same test stimuli that were primed for an exchange of initial consonants in the test condition, were also used in the baseline condition where these stimuli were preceded by 5 precursors that were *not* priming for an exchange error. In data sets 3–6 this baseline condition was omitted. In data sets 1 and 2 either a pair of real words or a pair of nonwords was elicited, in the data sets 3–6 nonwords were only sporadically elicited. There is an important difference between data sets 1, 2, 3, 5 and data sets 4 and 6 in the contrasts among the word-initial consonants. In data sets 1, 2, 3, and 5, elicited interaction was always between segments that differed in place or manner of articulation (one phonological feature, condition "pm1", which we consider a medium phonetic contrast) or in place as well as manner of articulation (two phonological features, condition "pm2", considered a strong phonetic contrast). The strength of contrast appears to be correlated with the relative number of speech sound errors against that contrast (Dell, 1986; Nooteboom, 1969; Nooteboom and Quené, 2017; Shattuck-Hufnagel and Klatt, 1979).

In data sets 5 and 6 the same test stimuli were used as in data sets 3 and 4, but two kinds of oppositions were added, viz. opposition between vowel sounds (condition "vowel") and opposition between voiced and voiceless consonants (one phonological feature, condition "voi"), thus doubling the number of test stimuli in data sets 5 and 6. However, we will ignore in the present paper all stimuli eliciting vowel exchanges (and their responses); the reason is that the difference in relative timing between initial consonants and nuclear vowel makes the vowel errors incomparable in their timing to the consonant errors. Although the voicing contrast involves a single phonological feature (as does the place contrast or manner contrast), it has been shown to be relatively weak for Dutch speakers (Van Alphen and McQueen, 2006; Van Alphen and Smits, 2004), and Dutch listeners perceive voicing of initial consonants less accurately than either place or manner (Cutler et al., 2004, p.3674); we will therefore regard the voicing contrast as a weak phonetic contrast. Note that the stimulus materials in our SLIP experiments was

Table 3
Response categories in the SLIP experiments; the example stimulus is "zaal boom" for which the exchange real-word error *baal zoom* was elicited. Note that categories 1, 2 and 3 jointly refer to elicited single errors, and that a full exchange of two initial consonants is regarded as a single error (cf. Shattuck-Hufnagel, 1983). Partial errors such as anticipations (*baal boom*) and perseverations (*zaal zoom*) are also regarded as elicited single errors.

Category	Label	Description	Example response
0	fluent correct	fluent and correct response, no error	*zaal boom*
1	early	elicited single error, repaired early	*baa… zaal boom*
2	late	elicited single error, repaired late	*baal zoom… zaal boom*
3	unrepaired	elicited single error, unrepaired	*baal zoom*
4	other	error other than elicited, including multiple errors	*boog baan*
5	hesitation	hesitation, omission	*well… eh…*

Table 4
Overview of response categories in the 6 data sets from SLIP experiments. Responses to all filler stimuli and to stimuli eliciting vowel exchanges have been excluded. "Other responses" (including non-elicited errors and multiple errors) and hesitations (including hesitations and omissions) were not further analyzed.

Data set	Response category				
	correct & fluent (0)	single elicited error (1, 2, 3)	other error(s) (4)	hesitation (5)	total responses
1	6510	238	390	206	7344
2	6788	116	292	148	7344
3	2850	201	228	113	3392
4	2873	211	263	45	3392
5	4592	511	451	110	5664
6	4895	526	432	99	5952
total	28,508	1803	2056	721	33,088

varied at the time in terms of phonological features of consonants (place, manner, voicing), whereas we now believe that the phonetic (articulatory, somatosensory, perceptual) contrast between the interfering speech sounds may be more relevant in speech production than their phonological representations.

For the present paper, earlier transcriptions of all recorded spoken responses were double-checked by the second author, and were categorized using the response categories shown in Table 3. The classification of errors into categories 1 (repaired early) and 2 (repaired late) will be explained and motivated below.

Table 4 provides a survey of some quantitative properties of the data sets that are included in the current investigation.

With single elicited speech errors constituting only 5.6 % of the responses, the SLIP technique is indeed very inefficient. Moreover, elicited errors were distributed very unequally over participants (Gini index 0.61), as illustrated in Fig. 1: 63 participants (15 %) did not produce any elicited error (light bar in Fig. 1), and most participants produced only a single elicited error. The top 77 (or 18 %) contributing participants produced half of the elicited errors (dark bars in Fig. 1), and the top 177 (or 41 %) participants contributed 1442 (or 80 %) of the elicited errors. The distribution of errors over stimuli is similarly unequal (Gini index 0.68). In the new analyses reported below, Bayesian methods will allow us to include both participants and stimulus items as random effects, thus taking these inequalities into proper account, even while the numbers of errors are low.

5. Results

5.1. Question 1

We have seen before that segmental speech errors can be detected by self-monitoring both before and after speech initiation has started. This raises the first question: what is the delay between "early" and "late" error-detection? If indeed very short error segments reflect self-monitoring "early" speech preparation and longer error segments reflect self-monitoring "late" speech preparation, then one would expect that the distribution of error-to-interruption intervals over all repaired speech errors would be bimodal. In our SLIP experiments, in virtually all responses the maximum number of segments is 6: viz. CVC CVC. In all completed responses we have taken the end of utterance as interruption.

Nooteboom (2005) measured such intervals in terms of the number of segments spoken after error onset and before interruption. He found a distribution that was clearly bimodal. Nooteboom and Quené (2017) measured such intervals in terms of log ms. They too found a distribution with two peaks, separated by nearly 500 ms. Here we have repeated the assessment of both distributions with more data, using multimodal analysis (Ameijeiras et al., 2019, 2021) in R (R Core Team, 2022). First, we inspect the distribution of the numbers of segments spoken between the error onset and the interruption, in single, elicited, repaired errors.

Fig. 2 suggests that there are two underlying distributions of number of segments, showing little overlap. We assume that all cases with segmental sequences lower than the separation value of 4.43 correspond to errors detected "early" and all cases with segmental sequences equal to or greater than the separation value correspond to errors detected "late". Of course, this classification of repaired speech errors will contain some misclassifications due to the slight overlap between the two underlying distributions, but given the relatively good separation, the number of misclassifications is statistically probably not very important. However, because the two underlying distributions do not appear to be normal, Gaussian, distributions, it is not easy to use these distributions to study further questions.

The other method of separating the two underlying distributions of repaired segmental errors is based on the time interval in log ms between error onset and interruption (again only for single, elicited, repaired errors). An uninformed mixture of Gaussian distributions (Scrucca et al., 2016) was fit, yielding two underlying Gaussians, as illustrated in Fig. 3. (Here 5 responses were ignored because of missing error-to-cutoff times, and 5 were ignored because of outlier values lower than 15 ms.)

One may note that the distribution of early detected errors, supposedly corresponding to errors detected during speech preparation,

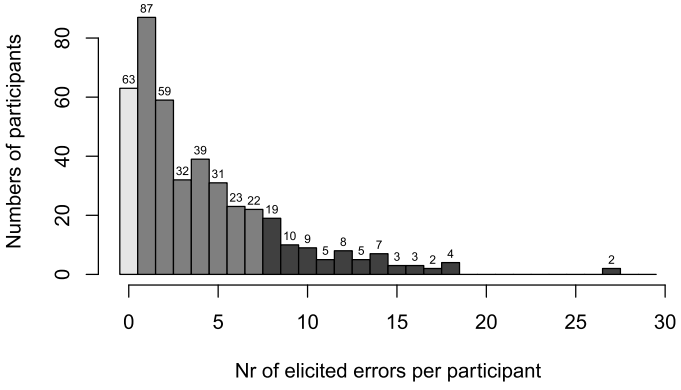


Fig. 1. Unequal distribution of the 1803 single elicited errors over 433 participants (light bar: 63 participants contributing no such errors; grey bars: relatively many participants contributing few such errors; dark bars: relatively few participants contributing many of such errors).

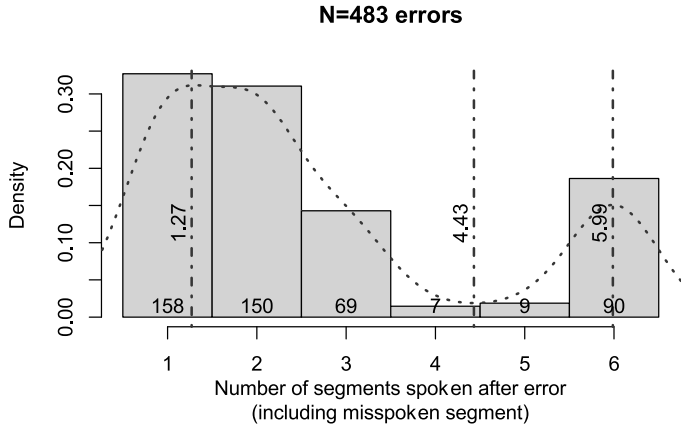


Fig. 2. Histogram (grey) and density (dotted) of the numbers of segments spoken between error onset and interruption, in 483 single, elicited, repaired errors. The boundary value is estimated by a multimodal test (Ameijeiras-Alonso et al., 2019).

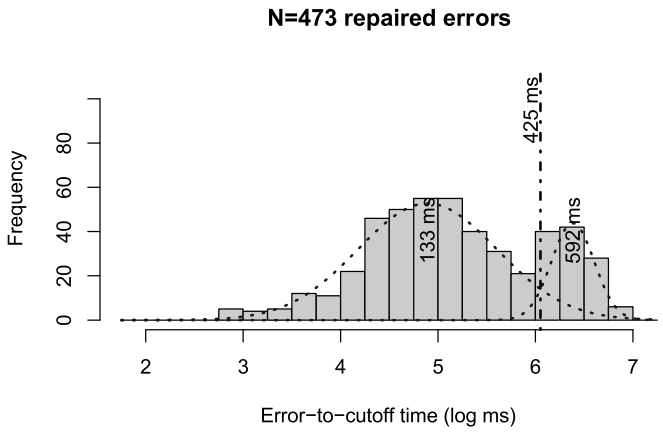


Fig. 3. Histogram of log-transformed durations in log ms of error-onset-to-interruption intervals, for 473 single, elicited, repaired errors. For completed responses the end of utterance counts as interruption. Dotted lines indicate the estimated distributions from an uninformed Gaussian mixture model. The vertical dashed line indicates the interpolated boundary value (at 6.05 log ms, corresponding to 425 ms) between the two estimated distributions. The peak values are given in ms.

tapers off at the lower side as if there are no or only very few further errors that were both detected and repaired before speech initiation. This suggests that such covertly detected and repaired errors (cf. Levelt et al., 1999) are very rare in our SLIP experiments (as also suggested by observing only 5 outlier values shorter than 15 ms). The upper side of the late distribution however seems to be cut off more sharply. This we regard as an artifact of the stimuli being used, which typically contained only 6 speech segments: CVC CVC. The distribution would look different if taken from errors in normal spontaneous speech. The right-censoring after completed responses (6 segments) may also explain why the distribution of "late" error-to-interruption is much narrower than that of "early" error-to-interruption times. Obviously, this is also an artifact.

After these preliminaries, the delay between "early" and "late" error detection can be estimated from the distance between the two peaks in the bimodal distributions in Figs. 2 and 3. This distance corresponds to 4.72 segments in Fig. 2 and to 459 ms in Fig. 3. The lower boundary of this delay, presumably being more informative of the delay between "early" and "late" speech preparation, may be estimated by comparing the lower tails of the two Gaussian distributions in Fig. 3: their 2.5 % percentile points correspond to 33 and 384 ms, respectively, yielding an estimated lower boundary of c. 350 ms for the delay between the detection of "early" and "late" errors. Presumably, the difference between these lower boundaries of "early" and "late" error detection time, corresponding to the earliest possible moments a speech sound error can be detected at each of the two stages of self-monitoring, reflects the temporal delay between "early" and ``late" preparation of speech sounds. We will come back to this in § 6.1 below.

5.2. Question 2

Why are some speech errors detected "early", others "late" and others again not at all? We have predicted in Section 2 that weak-contrast errors are less often detected and repaired than strong-contrast errors, and also that weak-contrast errors are less often detected "early" than strong-contrast errors. In comparing these odds of detection, we need to take into account that errors involving weak contrast occur more frequently than errors involving strong contrast (Dell, 1986; Nooteboom, 1969; Nooteboom and Quené, 2017; Shattuck-Hufnagel and Klatt, 1979). First, therefore, we verified whether this latter pattern also occurs in our SLIP experiments, by comparing the odds of an elicited error, and the odds of other errors, against the baseline of fluent and correct responses (see Table 4; N = 28,508 responses) using a Bayesian mixed-effects multinomial model. The population-level predictor was the strength of phonetic contrast (weak for voicing contrast (*voi*); medium for place or manner contrast (*pm1*); strong for place and manner contrast, *pm2*). In addition, random intercepts were added for participants and stimulus items, and the effects of phonetic contrast were allowed to vary within participants and between items ("random slopes"). This mixed-effects multinomial model was estimated using package *brms* in R (Bürkner, 2017, 2018) in R (R Core Team, 2022). The model was estimated in 4 independent chains of 3000 iterations (with 1000 warmup), using NUTS sampling. This yielded 8000 post-warmup iterations. For group-level ("random") estimates in the Bayesian models, we report the 95 % credibility interval (CrI) of the posterior distribution. For population-level ("fixed") estimates, we report the 95 % highest posterior density interval (HDI; Makowski et al., 2019; McElreath, 2020), which is the narrowest interval containing 95 % of the probability mass of the posterior distribution. If two model parameters have non-overlapping CrIs or HDIs, then we have good grounds to believe that those parameters are different.

The **group-level**("random") coefficients of this multinomial *odds-of-error* model of responses showed that between-item variation (in items' odds per category) was similar across conditions of phonetic contrast, with overlapping credibility intervals. Between-participant variation (in participants' odds of the elicited error) was higher in conditions eliciting errors in *voicing* as compared to the condition eliciting errors in *place and/or manner* (elicited errors: sd(*intercept*) 0.75 [0.63, 0.87], *pm2* +0.30 [0.04, 0.56], *voi* +0.46 [0.22, 0.68]; other errors: sd 0.75 [0.66, 0.86], *pm2* +0.18 [0.01, 0.39], *voi* +0.16 [0.01, 0.41]). Thus, participants are less similar in their propensity of voicing errors than in their propensity of errors involving place and/or manner—in agreement with voicing errors having been elicited in only two of the six SLIP experiments. The odds of an elicited error in *pm1* and in *voi* conditions were correlated between participants (r = −0.70 [−0.92, −0.42]); other between-participant correlations in odds of errors were not credibly different from zero.

The **population-level** coefficients of the first model capture the log odds of an error (against a fluent and correct response), broken down by phonetic contrast. The odds of an elicited error are highest if a voicing error is elicited (weak contrast *voi,* odds 588/3092; median of posterior log odds −1.88 [−2.08, −1.67]), considerably lower if a place-or-manner error is elicited (medium contrast *pm1*, odds 709/11,661, median −3.12 [−3.34, −2.91]), and lowest if a place-and-manner error is elicited (strong contrast *pm2*, odds 506/13,755, median −3.68 [−3.88, −3.47]), with non-overlapping 95 % HDIs. The odds of other (non-

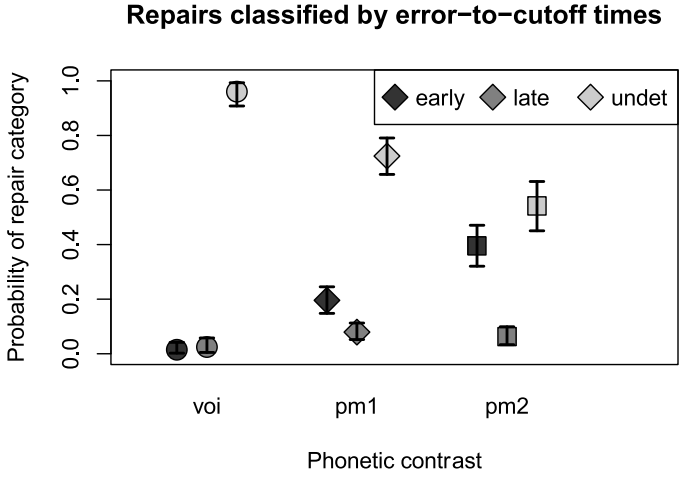


Fig. 4. Estimated probabilities of detection (and repair) categories based on error-to-cutoff time broken down by strength of phonetic contrast (*voi:* weak, *pm1:* medium, *pm2:* strong). Symbols are plotted at the median of the posterior distributions, and error bars denote their 95 % credibility intervals.

elicited) errors, not shown in Fig. 4, follow the same pattern, with medians at −2.63 [−2.87, −2.41], −2.81 [−2.99, −2.65] and −3.06 [−3.22, −2.91] respectively, but with overlapping 95 % HDIs.

In order to test the predicted effect of strength of contrast on detection-and-repair, our second (sequential) model zooms in on the *rates of early repair, of late repair and of non-repair* of the elicited errors only ($N = 1803$ observations). This model assumes a sequential, ordinal outcome variable: a late repair can only occur if there has been no early repair, and a non-repair can only occur if there has been neither an early nor a late repair (Bürkner and Vuorre, 2019). The random structure in this model was the same as in the first model above, and the strength of the phonetic contrast of the elicited error was once more included as a predictor. In addition, the model assumes category-specific effects of phonetic contrast (e.g., the effect of contrast *pm2* on the first threshold between early and late repair may differ from the effect of this same contrast on the second threshold between late repair and no repair, etc.), and it assumes unequal variances for the three phonetic contrasts. Details of the prior distributions and of the sampling, as well as summary measures of the posterior distributions, are provided in the Supplementary Materials. Fig. 4 summarizes the corresponding predicted rates in the three repair categories, for the three phonetic contrasts.

The **group-level** coefficients of this sequential "rates of repair category" model of elicited errors show that between-item variation (among items' rates per repair category) was higher both in the *pm2* condition (1.75 [0.57, 4.31]) and in the *voi* condition (1.54 [0.64, 2.96] than in the *pm1* condition (0.38 [0.22, 0.56]) where items were more similar. Between-participant variation (among participants' rates per repair category) was similar across conditions of phonetic contrast, with overlapping credibility intervals. Between-participants correlations in their repair categories across conditions were not credibly different from zero, except for the correlation between participants' random intercept and random slope of *pm2* (0.63 [0.10, 0.95]) indicating a possible floor effect.

The **population-level** coefficients of this sequential model are illustrated in Fig. 4 above. The rates of *early* repair of an elicited error are lowest if a voicing error was elicited (*voi*, 22/588), considerably higher if a place-or-manner error was elicited (*pm1*, 151/709), and highest if a place-and-manner error was elicited (*pm2*, 198/506), with non-overlapping 95 % credibility intervals. The rates of *late* repair are very low, and approximately the same across the phonetic contrasts (*voi* 27/588, *pm1* 55/709, *pm2* 30/506), with overlapping credibility intervals. Thus for late repairs there seems to be no effect of phonetic contrast; we will come back to this in the general discussion. The rates of *non*-repair of an elicited error are highest if a voicing error is elicited (*voi*, 539/588), considerably lower if a place-or-manner error is elicited (*pm1*, 503/709), and lowest if a place-and-manner error is elicited (*pm2*, 278/506), with non-overlapping 95 % credibility intervals of the posterior distributions.

Regarded differently, the prevalence of early repair over late repair increases with phonetic contrast: there is no such prevalence for voicing errors; the prevalence is strongest for repair of errors involving both place and manner contrasts, with an intermediate prevalence for errors involving either a place or a manner contrast. Thus elicited speech errors involving a voicing contrast remain mostly undetected and unrepaired, whereas of the elicited speech errors involving both place and manner over a third is detected and repaired early. These findings indicate that contrast strength is a major factor in predicting whether a particular single elicited segmental speech error is detected "early", "late", or not all. The pattern in our data also confirms that in Dutch the voicing contrast is relatively weak, as has also been found by Van Alphen and McQueen (2006) and Van Alphen and Smits (2004).

5.3. Question 3

As argued in §3 above, we hypothesize that—if a speech sound error has been made—the intended non-erroneous target is available as a repair candidate. Hence, after detection of a single elicited error, we predict firstly (a) that the most probable repair is the correct response candidate (as predicted by most theories of self-monitoring), and subsequently (b) that repairs after "early" detection are more often correct than those after "late" detection, and (c) that the interruption-to-repair times too are shorter after "early" than after "late" detection. These latter two predictions were confirmed by Nooteboom and Quené (2020); here we will re-test the same predictions using our expanded data set and more advanced analyses.

In order to test the predicted association between correctness of repair and the moment-of-interruption, we checked the $N = 469$ single, elicited consonant errors with an interpretable repair. For early-detected errors, the log odds of a repair being the correct response were log(343/28) = 2.51, while for late-detected errors the log odds were far lower at log(72/26) = 1.02. Firstly, these positive log odds confirm that overwhelmingly in SLIP experiments repairs are formed by the correct response candidates, thus confirming prediction (3a). If we disregard the random effects of participants and items, these aggregate numbers also suggest, secondly, that the odds of correct repair differ between early-detected and late-detected errors ($x^2(1) = 25.6$, $p < 0.001$); however, a more appropriate and more advanced Bayesian binomial mixed-effects model does not suggest such a difference in odds. In this latter model, random intercepts were added for participants and stimulus items, and the effects of the predictor varied across participants and items ("random slopes"), as in our previous models. Details of the modeling were the same as before, except that this model was estimated over 40,000 post-warmup iterations. The **population-level** log odds of correct repair were estimated at 3.31 for early-detected errors (with 95 % HDI [2.19, 5.05] including the observed aggregated odds of 2.51), and were estimated to be only slightly lower at 2.96 for late-detected errors (with larger 95 % HDI [0.76, 6.40] again including the observed aggregate odds of 1.02). The absence of evidence for a difference, according to this Bayesian mixed-effects binomial model, may well be due to the large **group-level** variability of the predicted effect of moment-of-interruption ("random slope") across participants ($s = 3.13$ [0.20, 9.18], $n = 238$) and across stimuli ($s = 2.41$ [0.14, 7.24], $n = 127$). Closer inspection of our corpus showed that only 2 out of 238 selected participants (those who had spoken any repair after any single elicited error) had made both correct and incorrect repairs after both early-detected and late-detected errors, with only 11 useful observations to assess repair patterns within these 2 participants. This scarcity of data obscures the predicted effect (3b) that is however visible in the aggregated numbers. In sum, the odds of correct repair (over other repairs) are consistent with decreasing activation during the delay between "early" and "late" error detection, but the statistical evidence is inconclusive.

From the difference between activation levels of the correct response candidates following "early" and "late" error detection we also predicted, thirdly, that the interval between interruption and repair is shorter after "early" than after "late" detection. Here, we need to consider that 30 (6 %) out of the 483 single, elicited, repaired errors were repaired immediately, with zero delay between interruption and repair. Therefore we analyzed interruption-to-repair times by means of a hurdle lognormal model (Heiss, 2022), which may be regarded as a combination of a logistic model (fitting the log odds of zero cutoff-to-repair time) and a log-normal model (fitting the above-zero cutoff-to-repair times). In this model, the random structure was the same as in previous models reported above. Here we only report the model including as a predictor the classification of the error as "early" or "late" based on its error-to-interruption time. (The parallel model, including the classification by number of segments spoken between error and interruption, is reported in the Supplementary Materials; the two models show highly similar outcomes). The predictor was also included as a group-level effect ("random slope") within participants and within items in both the hurdle part and the lognormal part of each model. The **population-level** coefficients of this model indicate that the log odds of a zero interruption-to-repair time are −4.94 [−7.92, −2.88]

for early-detected errors and −15.26 [−33.59, −4.30] or practically nil for late-detected errors; the posterior median of cutoff-to repair time after early-detected errors is 4.97 [4.90, 5.06] log ms or 144 [134, 158] ms, and after late-detected errors it is 5.69 [5.52, 5.86] log ms or 296 [250, 351] ms, with non-overlapping HDIs. (Group-level coefficients are reported in the Supplementary Materials).

According to our view of the repair process, these findings reflect the difference in activation level of the repair candidate caused by the long delay between self-monitoring "early" and "late" speech preparation. After early detection, but not after late detection, alternative candidates are sometimes still available for immediate repair. Interruption-to-repair times are considerably longer after late-detected errors than after early-detected errors, presumably because re-activation of repair candidates with decreased activation sometimes takes a considerable amount of extra time (cf. Seyfeddinipur et al., 2008; Tydgat et al., 2012).

In this Section 5 we have attempted to answer, mainly on the basis of data obtained in experiments on self-monitoring for segmental speech sound errors, three questions relating to the representation and production of speech sounds. We will continue with a discussion of our findings focusing on properties of the mental preparation of speech.

6. Discussion: speech preparation

Summarizing our findings, we found in our corpus of elicited speech errors and repairs (1) that there are two stages of self-monitoring with an average delay between "early" and "late" error detection by self-monitoring in the order of 460 ms. We also found (2) that a main factor in determining whether a segmental speech error is detected "early" or "late" or not at all, is strength of phonetic contrast between the intended and the intrusive consonant. As to repairs, we found (3) that the most frequent repair is the correct response, and that correct repairs seem to be more prevalent after "early" detection than after ``late" detection.

In the following section we will attempt to draw some conclusions relating to properties of speech preparation from our observations on self-monitoring.

6.1. The delay between "early" and "late" speech preparation

The distance between the two peaks in the bimodal distribution of intervals between error onset and interruption for repair (or end of utterance in completed responses) is assumed to be a measure of the delay between these two stages of self-monitoring. Under this assumption, the average delay is about 4 segments or 460 ms counted from the onset of the overt initial error consonant to interruption (cutoff), or to the end of utterance in full CVC CVC responses containing an error. However, using either way to express the delay (cf. Figs. 2 and 3) the distribution of error-to-interruption intervals is composed of (at least) two rather broad distributions of errors repaired "early" or "late", by multiple speakers under varying experimental conditions. Moreover, at least for the internally detected repaired errors, we are not actually looking at intervals between moment of error onset and interruption. We have no (direct) access to the error onsets in internal speech.

Let us have a closer look at Fig. 3 again. The lower boundary of the delay between "early" and "late" error detection may be estimated by comparing the lower tails of the two Gaussian distributions in Fig. 3: their 2.5 % percentile points correspond to 33 and 384 ms, respectively, yielding a lower boundary of 351 ms for the shortest delay between "early" and "late" error detection. Let us assume that the fastest reaction times to errors detected "early", i.e. in internal speech, are of the same order of magnitude as those to errors detected "late". With this assumption, the time between the most rapid detection of an internal error onset and speech initiation of that error sound is also in the order of 350 ms. This in turn suggests that this delay of 350 ms in some way corresponds to what happens between the preparation of speech sounds in internal speech and the preparation of the coordinated motor commands necessary for articulation is also in the order of 350 ms. Of course, this is considerably less than the time delay between "early" and "late" self-monitoring, as estimated from the distance between the two peaks of the distribution, which corresponds to 460 ms. The latter value of 460 ms difference probably is an underestimation of the delay under normal communication conditions, because of the artificial limitation in SLIP experiments of the utterances to 6 segments. (If the stimuli had been longer, then more segments of completed errors, requiring more speaking time, might have been spoken before interruption). Possibly, in continuous speech the average distance between ``early" and ``late" detection of segmental errors would be considerably greater than 4 segments or 460 ms. This could in principle be investigated by studying error-to-interruption intervals in repaired speech errors in continuous speech. Obviously, reacting to ``late" error detection on average takes more time than reacting to ``early" error detection. This does not necessarily reflect a delay in preparation of speech sounds, but more probably a delay in repairing caused by the weakened activation of repairs after ``late" as compared to ``early" error detection. It has been found that interruption may be postponed in the absence of a suitable repair candidate (Seyfeddinipur et al., 2008; Tydgat et al., 2012).

The time delay between the fastest "early" detections and the fastest "late" detections gives us a relevant estimate of the time delay between internal preparation and articulatory preparation of speech segments. If our analysis is more or less right, this estimated time delay, at least in our SLIP experiments, is in the order of 350 ms. This is a considerable amount of time. Apparently, transducing internal speech into (somatosensory targets necessary for) coordinated motor commands for articulatory gestures is a major operation. This suggests to us that speech sounds in internal speech are not represented in terms of articulation.

This is supported by evidence that the articulatory representation of speech is not yet programmed in internal speech: Nooteboom and Quené (2013) excised CV speech fragments from utterance initial segments in CVC CVC utterances, spoken in SLIP experiments (in fact, the two experiments from which data sets 1 and 2 in the current paper were taken). These word initial fragments were targeted for eliciting interactive speech errors. Each fragment for each condition from the same speaker in the same experiment, stemmed from undetected (unrepaired) speech errors, or from "early" detected repaired speech errors, or from "late" detected repaired elicited speech errors, or from the corresponding fluent and correct productions as a control condition. The idea was to find out whether ubiquitous blending of articulatory gestures as demonstrated by Goldstein et al. (2007) and McMillan and Corley (2010), which often cannot be detected auditorily (cf. Pouplier and Goldstein, 2005), would nevertheless affect speech perception as assessed by reaction times in a phoneme identification experiment. Our prediction was that error segments from both "early" detected and "late" detected errors, and also from undetected errors, would all show longer average identification times than the control condition, due to (unknown acoustic-phonetic effects of) articulatory blending. This would support the conclusions by Goldstein et al. (2007) and McMillan and Corley (2010). This is, however, not what was found. Instead, the only condition with clearly longer reaction times than in the control condition was formed by the "late" detected errors. "early" detected errors did not lead to longer reaction times, undetected errors were in between, possibly because these form a mix of errors made internally and errors only made on the level of somatosensory targets or motor commands. In retrospect (at the time we published these results we had no convincing interpretation), we conclude from our earlier findings that articulatory blending is not yet programmed in internal speech, and that speech sounds are indeed coded in internal speech as targets in auditory perceptual space, as proposed by Guenther (2016).

Results obtained by Goldrick and colleagues (Alderete et al., 2021; Goldrick and Blumstein, 2006; Goldrick et al., 2016) do not seem to agree in all respects with our results. Goldrick and Blumstein (2006) demonstrated cascading of information from a previous level of speech preparation to articulation by showing that errors in elicited tongue

twisters contain subtle traces of the correct versions. However, the discrepancy vanishes if we assume that their method of eliciting tongue twisters, very similar tot the method by Goldstein et al. (2007), does not elicit errors in internal speech but only elicits errors on the level of motor commands for articulatory gestures. However, there is one indication that "early" error detection (in internal speech) and "late" error detection (presumably, under time pressure, on the level of motor commands for articulation), behave differently: whereas we find a strong effect of phonetic contrast on the frequencies of "early" detection and non-detection, there is no effect on the frequency of "late" detection. Phonetic contrasts were varied in abstract features (voice; place-or-manner; place-and-manner), and *not* in articulatory gestures. Consequently, contrasts differ in their abstract features involved, but the contrasts may be more equal (or even neutralized) in terms of articulatory gestures. For example, the medium contrast [p]~[t] (place-only) and the strong contrast [b]~[z] (place-and-manner) both differ in two articulatory gestures (absent in one and present in the other consonant). More generally, the contrasts involved in our corpus tend to differ more strongly in their abstract features than in their associated articulatory gestures. This might explain the absence of an effect of (abstract) phonetic contrast in the frequencies of "late" detections; further research is necessary to corroborate this tentative explanation.

We also point out that if there is somatosensory (tactile and/or proprioceptive) feedback from articulation in self-monitoring for speech errors, this means that speech has to be represented not only in terms of auditory targets and articulatory gestures, but also in terms of somatosensory signals from the articulators, as proposed by Guenther (2016). Feedback from articulation in addition to feedback via the self-produced acoustic waveform and audition, has also been proposed by Hickok (2012), Lackner and Tuller (1979), Nooteboom and Quené (2017) and Pickering and Garrod (2013). The kinds of activation of speech sounds during speech preparation can in principle be further investigated by neuro-imaging research.

6.2. Simultaneously activated response candidates compete for the same slot

In this paper we have also supported the proposal by Nozari et al. (2011; see also Nozari and Pinet, 2020) that during speech preparation, simultaneously activated candidate responses may compete for the same slot. This is obviously the case in SLIP experiments, where the conflict between competing "speech plans" is set up in the experimental technique, but such competition may also occur frequently during speech preparation in other conditions (Nooteboom and Quené, 2019).

In principle, competition can exist at different levels of speech preparation, and between or within units of different sizes, and resulting "blends" may even become lexicalized. We can have competition between lexical phrasal items such as verbal idioms, leading to phrasal blends: I'm going to give him a taste of my mind! (a piece of my mind plus a taste of his own medicine), between lexical units: *brunch* (from *breakfast* and *lunch*), and also of phoneme-size speech sounds, leading to added or suppressed articulatory gestures as attested by Goldstein et al. (2007, see §2 above).

Interestingly, it is not always easy to know what the units involved in particular interactive segmental speech errors are. When /p/ turns into /k/ in the speech error *kaf piep* instead of *paf kiep*, this may be regarded as an error on the segmental level. But it is relevant that the exchanged segments are both in initial position, in lexical word forms. In earlier work we have shown that initial consonants mostly interact with initial consonants, medial consonants with medial consonants, and final consonants with final consonants (Nooteboom and Quené, 2015). It would be fair to say that in *kaf* for *paf* a lexical word form is misspoken by replacing a single speech sound of the lexical form by another single speech sound. The relevance of the word form as the unit that is misspoken and repaired, is confirmed in our study of repairs: Repairs hardly ever consist of single speech sounds, or single syllables. Virtually all repairs consist of at least the misspoken word form. In our SLIP material most repairs consist of complete CVC CVC word pairs.

On the lexical level of speech preparation, where units from the mental lexicon are activated together with the information necessary for the spell-out in terms of speech sounds, conflicting units competing for the same slot, of necessity are morphemes or words. During the spell-out, creating candidate word forms in terms of a sequence of speech sounds, conflicting units competing for the same position often are phoneme-size speech sounds labelled for the position within the lexical form. Much less frequent are speech errors where meaningless combinations of speech sounds, such as VC or CV, or consonant clusters behave as units. As, for example, observed by Fromkin (1973) and other students of speech errors, there are also speech errors where sub-phonemic features take each other's places, as in *mang the mail* for *bang the nail* (Fromkin, 1973). Although such errors are much less frequent than whole-segment exchanges, Fromkin (1973) gives a list of 55 such cases in English.

Our finding that repairs are often formed by correct candidate responses (§5.3), is in itself not very surprising, as this has often been observed. However, there does not seem to be general agreement on the mechanism underlying repair. We explicitly propose that, especially after "early" detection, a repair may stem from a competing (and often correct) candidate word form that is still active. This proposal is in line with a suggestion already made by Blackmer and Mitton (1991). The short interruption-to-repair times in our data (with posterior median 144 ms after "early" detection) suggest that indeed such repairs are drawn from still active word candidates. Of course, in case of prolonged interruption-to-repair times, this prolongation may have been caused by necessary re-activation of the repair response. One may recall that in our SLIP experiments interruption-to-repair times include many cases in which we counted the offsets of completed CVC CVC utterances as interruptions. In those cases prolonging the interruption-to-repair times takes the place of postponement of interruption as described by Seyfeddinipur et al. (2008) and Tydgat et al. (2012). These authors proposed that, in cases in which no repair is immediately available, interruption may be postponed until a repair is available. A repair may become available either by re-activating a weakly activated response candidate, or in extreme cases, by creating a new response candidate.

7. Conclusions

Our re-analysis of sets of responses obtained in earlier-published reports on SLIP experiment leads us to the following main conclusions: (1) Preparation of speech following the activation of lexical items, has at least two different stages. The transduction of planned speech sounds from the "early" to the "late" stage is a major, time-consuming operation. (2) We propose that at the "early" stage speech sounds are represented as phoneme-size units close to targets in auditory perceptual space and at the "late" stage as somatosensory targets activating coordinated motor commands necessary for articulation. Speech sounds are also represented in terms of somatosensory features, necessary for feedback from articulation to speech planning. (3) At the lexical level of speech preparation, for each slot in the sequence of lexical items, there may be conflict between simultaneously activated lexical items competing for the same slot. This conflict is carried over to the spell-out of lexical items in terms of speech sounds. At both the "early" and the "late" stage of speech preparation, the competing units are candidate word-like responses. A response with the next-highest activation remains active during and even after articulation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Supplementary materials (data and analyses) are available at https://osf.io/79ynw/.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2024.103043.

References

Alderete, J., Baese-Berk, M., Leung, K., Goldrick, M., 2021. Cascading activation in phonological planning and articulation: evidence from spontaneous speech errors. Cognition 210, 1–5.

Ameijeiras-Alonso, J., Crujeiras, R.M., Rodríguez-Casal, A., 2019. Mode testing, critical bandwidth and excess mass. *TEST* 28, 900–919. https://doi.org/10.1007/s11749-018-0611-5.

Ameijeiras-Alonso, J., Crujeiras, R.M., Rodríguez-Casal, A., 2021. multimode: an R Package for Mode Assessment. J. Stat. Softw. 97 (9), 1–32. https://doi.org/10.18637/jss.v097.i09.

Baars, B.J., Motley, M.T., 1974. Spoonerisms: experimental elicitation of human speech errors. J. Suppl. Abst. Serv.: Cat. Sel. Documents Psychol. 3, 28–47.

Blackmer, E.R., Mitton, J.L., 1991. Theories of monitoring and the timing of repairs in spontaneous speech. Cognition 39, 173–194.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer (Version 6.0.19) [Computer program]. Available at <http://www.praat.org/>.

Bohland, J.W., Bullock, D., Guenther, F.H., 2010. Neural representations and mechanisms for the performance of simple speech sequences. J. Cogn. Neurosci. 22 (7), 1504–1529.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108 (3), 624–652.

Browman, C.P., Goldstein, L., 1992. Articulatory phonology: an overview. Phonetica 6 (2), 201–225.

Bürkner, P.-C., 2017. brms: An R package for Bayesian multilevel models using Stan. J. Stat. Softw. 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C., 2018. Advanced Bayesian multilevel modeling with the R package brms. R. J. 10 (1), 395–411. https://doi.org/10.32614/RJ-2018-017.

Bürkner, P.-C., Vuorre, M., 2019. Ordinal Regression Models in Psychology: a Tutorial. Adv. Methods Pract. Psychol. Sci. 2 (1), 77–101. https://doi.org/10.1177/2515245918823199.

Cutler, A., Weber, A., Smits, R., Cooper, N., 2004. Patterns of English phoneme confusions by native and non-native listeners. J. Acoust. Soc. Am. 116 (6), 3668–3678.

Dell, G.S., 1986. A spreading-activation theory of retrieval in sentence production. Psychol. Rev. 93, 283–321.

Gauvin, H.S., Hartsuiker, R.J., 2020. Towards a new model of verbal monitoring. J. Cogn. 3 (1), 1–37.

Goldrick, M., Blumstein, S.E., 2006. Cascading of activation from phonological planning to articulatory processes. J. Lang. Cogn. Process. 21, 649–683.

Goldrick, M., Keshet, J., Gustafson, E., Heller, J., Needle, J., 2016. Cognition 149, 31–39.

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

Guenther, F.H., 2016. Neural Control of Speech. MIT Press, Cambridge Ma.

Hartsuiker, R.J., Corley, M., Martensen, H., 2005. The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: related Beply to Baars, Motley, and MacKay (1975). J. Mem. Lang. 52, 58–70.

Hartsuiker, R.J., Kolk, H.H.J., 2001. Error monitoring in speech production: a computational test of the perceptual loop theory. Cogn. Psychol. 42, 113–157.

Heiss, A. (2022). *A guide to modeling outcomes that have lots of zeros with Bayesian hurdle lognormal and hurdle Gaussian regression models.* Blog available at https://www.andrewheiss.com/blog/2022/05/09/hurdle-lognormal-gaussian-brms/#hurdle-model-with-additional-terms.

Hickok, G., 2012. Computational neuroanatomy of speech production. Nat. Rev. Neurosci. 13, 135–145.

Lackner, J.R., Tuller, B.H, 1979. Role of efference monitoring in the detection of self-produced speech errors. In: Cooper, W.E., Walker, E.C.T. (Eds.), Sentence Processing. Erlbaum, Hillsdale, N.J., pp. 281–294

Laver, J.D.M., 1973. The detection and correction of slips of the tongue. In: Fromkin, V. A. (Ed.), Speech Errors as Linguistic Evidence. Mouton, The Hague, pp. 132–143.

Levelt, W.J.M., 1983. Monitoring and self-repair in speech. Cognition 14, 41–104.

Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. The MIT Press, Cambridge, Massachusetts.

Levelt, W.J.M., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Behav. Brain Sci. 22, 1–75.

MacKay, D.G., 1987. The Organization of Perception and Action: A theory for Language and Other Cognitive Skills. Springer-Verlag, Berlin.

McElreath, R., 2020. Statistical Rethinking: A Bayesian Course with Examples in R and Stan, second ed. CRC Press, Boca Raton.

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

Makowski, D., Ben-Shachar, M.S., Lüdecke, D., 2019. bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J. Open Source Softw. 4 (40), 1541. https://doi.org/10.21105/joss.01541.

Nooteboom, S.G., et al., 1969. The tongue slips into patterns. In: Sciarone, A., et al. (Eds.), Nomen, Leyden Studies in Linguistics and Phonetics. Mouton, The Hague, pp. 114–132. Also in V.A. Fromkin (ed.), Speech Errors as Linguistic Evidence, pp. 144–156, Mouton, The Hague (1973).

Nooteboom, S.G., 2005. Lexical bias revisited: detecting, rejecting and repairing speech errors in inner speech. Speech Commun. 47, 43–58.

Nooteboom, S.G., Quené, H., 2008. Self-monitoring and feedback: a new attempt to find the main cause of lexical bias in phonological speech errors. J. Mem. Lang. 58, 837–861.

Nooteboom, S.G., Quené, H., 2013. Parallels between self-monitoring for speech errors and identification of the misspoken segments. J. Mem. Lang. 69, 417–428.

Nooteboom, S.G., Quené, H., 2015. Heft lemisphere: exchanges predominate in segmental speech errors. J. Mem. Lang. 68, 26–38.

Nooteboom, S.G., Quené, H., 2017. Self-monitoring for speech errors: two-stage detection and repair with and without auditory feedback. J. Mem. Lang. 95, 19–35.

Nooteboom, S.G., Quené, 2019. Temporal aspects of self-monitoring for speech errors. J. Mem. Lang. 105, 43–59.

Nooteboom, S.G., Quené, H., 2020. Repairing speech errors: competition as a source of repairs. J. Mem. Lang. 111, 104069 https://doi.org/10.1016/j.jml.2019.104069.

Nozari, N., 2020. A comprehension- or production-based monitor? Response to Roelofs. J. Cogn. 3 (1), 1–21, 19.

Nozari, N., Dell, G., Schwartz, M., 2011. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. Cogn. Psychol. 63, 1–33.

Nozari, N., Pinet, S., 2020. A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. J. Neurolinguist. 53 (3), 100875.

Oppenheim, G.M., Dell, G.S., 2008. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. Cognition 106 (1), 526–537.

Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behav. Brain Sci. 36 (4), 329–347.

Postma, A., 2000. Detection of errors during speech production: A review of speech monitoring models. Cognition 77 (2), 97–132.

Pouplier, M., Goldstein, L., 2005. Asymmetries in the perception of speech production errors. J. Phon. 33, 47–75.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

Roelofs, A., 2020a. Self-monitoring in speaking: in defense of a comprehension-based account. J. Cogn. 3 (1), 1–13. *18*.

Roelofs, A., 2020b. On (correctly) representing comprehension-based monitoring in speaking: rejoinder to Nozari (2020). J. Cogn. 3 (1), 1–7, 20.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R. J. 8 (1), 289–317.

Seyfeddinipur, M., Kita, S., Indefrey, P., 2008. How speakers interrupt themselves in managing problems in speaking: evidence from self-repairs. Cognition 108 (3), 837–842.

Shattuck-Hufnagel, S., 1983. Sublexical units and suprasegmental structure in speech production planning. In: MacNeilage, P.F. (Ed.), The Production of Speech. Springer, New York, Heidelberg, pp. 109–136.

Shattuck-Hufnagel, S.R., Klatt, D.H., 1979. The limited use of distinctive features and markedness in speech production: evidence from speech error data. J. Verbal Learn. Verbal Behav. 18, 41–55.

Tourville, J.A., Guenther, F.H., 2011. The DIVA model: a neural theory of speech acquisition and production. Lang. Cogn. Process. 26 (7), 952–981.

Tydgat, I., Diependaele, K., Hartsuiker, R.J., Pickering, M.J., 2012. How lingering representations of abandoned context words affect speech production. Acta Psychol. 140, 189–229.

Van Alphen, P.M., McQueen, J.M., 2006. The effect of voice onset time differences on lexical access in Dutch. J. Exp. Psychol. 32 (1), 178–196.

Van Alphen, P.M., Smits, R., 2004. Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. J. Phon. 32 (4), 455–491.

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111 (4), 931–959.