Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

Parallels between self-monitoring for speech errors and identification of the misspoken segments

Sieb G. Nooteboom [*], Hugo Quené

Utrecht University, Utrecht Institute of Linguistics OTS, Trans 10, 3512 JK Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 17 December 2012
revision received 19 April 2013
Available online 22 May 2013

Keywords:
Speech errors
Self-monitoring
Speech perception
Identification
Reaction times

ABSTRACT

This paper investigates self-monitoring for speech errors by means of consonant identification in speech fragments excised from speech errors and their correct controls, as obtained in earlier experiments eliciting spoonerisms. Upon elicitation, segmental speech errors had been either not detected, or early detected or late detected and repaired by the speakers. Results show that misidentifications are rare but more frequent for speech errors than for control fragments. Early detected errors have fewer misidentifications than late detected errors. Reaction times for correct identifications betray effects of varying perceptual ambiguity. Early detected errors result in reaction times that are even faster than those of correct controls, while late detected errors have the longest reaction times. We speculate that in early detected errors speech is initiated before conflict with the correct target arises, and that in both early and late detected errors conflict between competing segments has led to detection.

Introduction

Speakers occasionally produce erroneous speech sounds. Does the speech sound resulting from an error constitute a categorically different speech sound, or is it a blend of competing speech segments? Speakers may also correct their speech errors. Does self-monitoring involve inspecting production processes during speech preparation or does it involve inspecting the end products of production processes by employing speech perception? In order to answer these questions, this paper describes an attempt to investigate aspects of speakers' self-monitoring for speech errors in an indirect way. To that end we had listeners identify speech fragments containing segmental speech errors and fragments containing corresponding correctly spoken control fragments. The errors had either been not detected (*bood geer*), or early detected and repaired (*boo...good beer*) or late detected and repaired (*bood geer...good beer*) by the original speakers. Listeners' error rates and reaction times in identifying these segments, taken from various types of speech errors, may provide answers to the questions above.

For a survey of models of self-monitoring the reader is referred to Postma (2000) and Nozari, Dell, and Schwartz (2011). In the course of this paper we will mainly focus on the differences between perception-based monitoring as exemplified in the perceptual loop theory of self-monitoring by Levelt (1989) and Levelt, Roelofs, and Meyer (1999) on the one hand and conflict-based monitoring as proposed by Nozari et al. (2011) on the other. We assume that from the perceptual loop theory one may infer that self-monitoring employs perceptual properties of speech sounds in error detection. We also assume that self-monitoring mainly employs perceptual comparison between error form and correct target form, as suggested by Nooteboom (2005a, 2005b) and Nooteboom and Quené (2008). It should be noted that the latter position has been criticized by McMillan and Corley (2010). They have

[*] Corresponding author. Address: Cor Ruyslaan 20, 3584 gd Utrecht, The Netherlands.
E-mail address: S.G.Nooteboom@uu.nl (S.G. Nooteboom).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.04.006

difficulty to see how such a comparison between error and correct target could fit in the cascading framework they propose, in which competing segments can be simultaneously activated and can activate simultaneous conflicting articulatory gestures. In that framework segmental errors are not all-or-none. This would complicate comparison between intended targets and error forms. McMillan and Corley also wonder why, if the correct target is available for comparison with the target form, an error form was generated in the first place. However, these objections notwithstanding, if self-monitoring would employ comparison between error form and target form, then one would predict that the probability of error detection increases with perceptual distance between error form and target form. This is different for conflict-based monitoring as proposed by Nozari et al. (2011).

Nozari et al. (2011) reject perception-based monitoring because of the often reported double dissociation between speech error detection and speech perception in aphasic patients (e.g. Butterworth & Howard, 1987; Liss, 1998; Marshall, Rapaport, & Garcia-Bunuel, 1985; Marshall, Robson, Pring, & Chiat, 1998). Their conflict-based monitor for speech errors is computationally implemented in the two-stage word production model described by Dell (1986), it is production-based, and monitors for conflict of activation between simultaneously activated units during speech preparation. This proposed self-monitoring system would fit in well with the cascading framework proposed by McMillan and Corley (2010). These authors suggest that the conflict among multiple simultaneously active segments, competing for the same slot in inner speech, may cascade down to articulation. This would then result in articulatory blending. The theory by Nozari et al. predicts that (in normal speakers) the probability of error detection increases with increasing amount of conflict between simultaneously activated units. Combining the conflict-based theory of Nozari et al. (2011) and the cascading of activation proposed by McMillan and Corley (2010), we see that according to the conflict-based theory of monitoring the probability of error detection increases with amount of articulatory blending. This is interesting because whereas it is difficult to measure the amount of conflict of activation (but see Botvinick, Braver, Barch, Carter, & Cohen, 2001; Yeung, Botvinick, & Cohen, 2004), we can in principle assess the amount of articulatory blending by means of a perception experiment, as we will report below.

Most speech errors are errors against single speech segments; roughly half of these segmental speech errors are detected and repaired by the speakers (cf. Nooteboom, 1980, 2005a; Nooteboom & Quené, 2008). Until not very long ago people studying segmental speech errors seemed to work from the assumption that most segmental speech errors are categorical in nature, that is that they consisted of the substitution, deletion or addition of complete segments. This despite the fact that in the last few decades it has been shown repeatedly that blends of simultaneously pronounced speech sounds are not infrequent. For example, Mowrey and MacKay (1990) demonstrated that segmental errors of speech elicited in the laboratory, every now and then contain electromyographic evidence of simultaneous competing segments. In an acoustic study of elicited confusions between /s/ and /z/, Frisch and Wright (2002) found that both categorical and gradient errors occurred, although categorical errors were more frequent than one would expect if they were just extreme examples of gradient voicing errors. Goldrick and Blumstein (2006), focusing on voice onset time in voiced and voiceless consonants, also found that in elicited segmental speech errors acoustic traces of the competing segments can be found. From these studies it seems apparent that many speech errors are gradient and not categorical, although the results still do not exclude the possibility that the majority of speech errors is categorical. More recently it has been demonstrated that (at least in a particular experimental setting) most segmental speech errors are not categorical errors but rather blends of competing articulatory gestures (Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; Pouplier, 2007; McMillan & Corley, 2010). It is at this stage not clear what causes the discrepancy between the earlier studies and the more recent studies, but all these studies agree that gradient segmental speech errors are frequent. The fact that, despite the relative frequency of articulatory blending, in the past canonical speech errors were assumed to be categorical instead of gradient may be attributed to perceptual illusions during transcription (Pouplier & Goldstein, 2005): Our perception is categorical also when the perceived produced speech segments are not.

If most speech errors indeed are articulatory blends of competing segments, this is likely to have consequences both for detecting such speech errors in self-monitoring and for the perceptual properties of these errors. In this paper we present an experiment exploring parallels between self-monitoring for segmental speech errors and perceptual identification of the misspoken segments, and testing some hypotheses stemming from the supposed blended origin of segmental speech errors. It should be noted that the prevalence among segmental speech errors of articulatory blends supports a cascading model of speech preparation in which segments may compete for the same slot in the speech plan (McMillan & Corley, 2010). This, in turn, makes the proposal by Nozari et al. (2011) of conflict-based monitoring for speech errors seem realistic.

The basic idea underlying this paper is the following. If Goldstein, Pouplier, Chen, Saltzman, and Byrd (2007) and McMillan and Corley (2010) are right, then speech segments resulting from segmental errors of speech often must carry the acoustic consequences of the articulatory blending of speech sounds. These acoustic consequences of articulatory blending must in turn have perceptual consequences, even if very often these consequences are not reflected in auditory transcription (cf. Pouplier & Goldstein, 2005; McMillan, 2008). If we excise speech fragments containing the erroneous segments from elicited speech errors and offer these speech fragments, together with speech fragments excised from correct controls (no speech errors), in a simple speech segment identification experiment, then the perceptual consequences of the assumed articulatory blending may become apparent in two dependent measures, viz. frequency of misidentifications and reaction times. This prediction is based on the assumption that articulatory blending has acoustic consequences (cf. Goldrick & Blumstein, 2006), and that these negatively affect perceptual clarity. Perceptually ambiguous or unclear segments will lead to more misidentifications and to longer reaction times than perceptually unambiguous and clear segments (for the effect of response conflict caused by perceptual ambiguity on reaction times see Botvinick et al., 2001; Szmalec et al., 2008). If indeed most segmental speech errors contain traces of articulatory blending, and most correctly spoken segments do not, then we can predict that speech fragments excised from error segments are more often misidentified and on average have longer reaction times than speech fragments from correct control segments (**prediction 1**).

The speech errors and their correct controls were taken from past experiments in which many hundreds of segmental speech errors were elicited in Dutch with the so-called SLIP technique, as described in Nooteboom and Quené (2008; for the SLIP technique see Baars & Motley, 1974; Baars, Motley, & MacKay, 1975). In these experiments all elicited erroneous and correct utterances consist of two CVC monosyllables, as in the English example *good beer*. Of each elicited utterance we know whether it was correct or a speech error and if a speech error whether or not it was detected and repaired by the speaker. From a perception-based theory of self-monitoring, one expects that error detection depends on perceptual properties of the segments concerned. Nooteboom and Quené (2008) have proposed that perception-based self-monitoring mainly depends on perceptual comparison between error and correct target form. If so, one expects that the probability of error detection increases with perceptual distance between error and correct target. Perceptual distance between error and target will be greater when the error is categorical than when the error is gradient. From this reasoning we predict that detected errors will be perceptually clearer than undetected errors and therefore will be less often misidentified and will on average have shorter reaction times than undetected errors (**prediction 2**). It should be noted that if one assumes that self-monitoring is not perception-based but rather, as suggested by Nozari et al. (2011) conflict-based, one makes the opposite prediction. This is so, because conflict between segments competing for the same slot in speech preparation will cascade to articulation and cause articulatory blending and as a result leads to some measure of perceptual ambiguity or unclarity. Therefore, from a conflict-based theory of self-monitoring one predicts that detected errors are more often misidentified and on average have longer reaction times than undetected errors.

If segmental speech errors had been detected and repaired we also know whether detection was early, in inner speech, or late, where detection could have been during articulation (cf. Postma, 2000) or in overt speech (cf. Nooteboom, 2005a; Nooteboom, 2005b; Hartsuiker, Kolk, & Martensen, 2005; Huettig & Hartsuiker, 2010). Early detected speech errors are those that are interrupted rapidly after speech initiation, such as *boo…good beer*. We know that such speech errors are detected by the speaker in inner speech and not in overt speech because in early interruptions the speech fragment virtually always is shorter than a humanly possible reaction time (Blackmer & Mitton, 1991; Nooteboom, 2005a; Nooteboom, 2005b; Nooteboom & Quené, 2008). It is also noteworthy that early interrupted speech errors are practically always followed by a repair, demonstrating that interruption is indeed caused by error detection. The offset-to-repair interval is often in the order of 0 ms, showing that not only interruption but also detection was planned before speech initiation (Blackmer & Mitton, 1991; Nooteboom, 2005b). Nooteboom and Quené (2008) found that response times for early interrupted segmental errors are significantly shorter than those for other speech errors. They proposed that early interruptions result from too hasty speech initiation, before self-monitoring has had a chance to detect the error in inner speech. When self-monitoring catches up a moment later, speech is interrupted. Late detected speech errors are defined here as those elicited speech errors that are repaired only after the whole erroneous utterance, consisting of two monosyllables, has been uttered. One might argue that late detected errors may also have been detected in inner speech but that the reaction to this detection has been delayed, for whatever reason. However, it is generally believed that late detected speech errors are not detected in inner speech but rather in overt speech (Cf. Nooteboom, 2005a; Nooteboom, 2005b; Hartsuiker & Kolk, 2001; Hartsuiker et al., 2005; Huettig & Hartsuiker, 2010). One can argue that from a perception-based theory of monitoring for speech errors, and notably from the assumption that monitoring is based on perceptual comparison between error form and correct target form, it is expected that early detected errors are perceptually clearer than late detected errors. This is so, because early detected errors would be detected early precisely because they are perceptually clear, and late detected errors would be detected late precisely because they are perceptually somewhat less clear. Thus from the perspective of a perception-based monitor for speech errors one predicts that early detected errors are less often misidentified and on average have shorter reaction times than late detected errors (**prediction 3).** This is different from the perspective of conflict-based monitoring: Early detected errors would be detected early precisely because the conflict between competing segments is relatively great, and therefore articulatory blending would be relatively strong, and this would cause relatively serious perceptual unclarity that could become apparent in a subsequent identification task, whereas late detected errors would be detected late precisely because conflict is somewhat less than in early detected errors and therefore articulatory blending and perceptual unclarity would be somewhat less serious. From this one predicts that early detected errors would be more often misidentified and would on average have longer reaction times than late detected errors. So again we see that perception-based self-monitoring and conflict-based self-monitoring lead to opposite predictions.

In the course of this introduction we have made the following predictions, derived from perception-based self-monitoring for speech errors:

1. Speech fragments excised from segmental speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from correct controls.
2. Speech fragments excised from undetected speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from detected speech errors. However, it may be noted that conflict-based monitoring makes the opposite prediction.
3. Speech fragments excised from late detected speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from early detected speech errors. But again, it may be noted that the opposite prediction can be derived from conflict-based monitoring.

Table 1
Breakdown of the stimuli used in the identification experiment, according to whether the elicited spoonerisms were completed or interrupted and to whether the speaker had or had not detected and repaired the speech error. Controls were always completed. In italics are given fictitious examples of excised fragments plus, between brackets, the remainders of the complete responses from which these fragments were excised.

Category	Detection	Number
Errors	Undetected *boo(d geer)*	158
Errors	Early detected *boo…(good beer)*	80
Errors	late detected *boo(d geer…good beer)*	53
Errors	**Total**	**291**
Correct controls	**Total**	**291**

Method

Stimulus material

We first selected speech errors made in the test condition of two experiments (done with Dutch participants in the Dutch language), described by Nooteboom and Quené (2008). Both experiments elicited initial consonant exchanges in pairs of CVC words. We selected all speech errors in which the initial consonant of the first word of the test word pair was replaced by the initial consonant of the second word of the test word pair (thus both *boo…good beer* and *bood geer…good beer* would have been selected). This gave us a set of 453 speech errors. However, not all of these errors had a correctly spoken control due to many omissions, hesitations and other kinds of errors in the base-line condition. The base-line condition in these experiments had the same to-be-spoken word pairs as the test condition, but the preceding word pairs, to be read silently, did not trigger an exchange of the two initial consonants. These preceding word pairs were phonologically not related to the to-be-spoken word pairs. This base-line condition gave us in principle the opportunity to find a correctly spoken utterance for each successfully elicited speech error. We removed all speech errors that for some reason or other did not have a correctly spoken control. This left us with 291 successfully elicited word pair initial speech errors, each with a correctly spoken control. Using Praat (Boersma & Weenink, 2009) each of the misspoken and each of the corresponding correctly spoken consonantal segments was excised from the original spoken word form, beginning immediately before sound onset and ending 40 ms after the vowel onset. This served the purpose of removing the lexical context, and also made it possible to include many of the cases in which the speech error was early interrupted, as in *good beer* > *boo…good beer*. In the latter case, the speech error obviously had been detected and repaired by the speaker in inner speech. The relevant breakdown of the stimuli is given in Table 1.

For all speech fragments the origin was coded as “correct control”, “undetected speech error”, “early detected speech error”, or “late detected speech error”. “Undetected speech errors” are those that were not interrupted and not repaired. “Early detected speech errors” are those in which only the initial CV of the error form was spoken, with at least 40 ms of the vowel. “Late detected speech errors” are those that were not interrupted but only repaired after the full error utterance was spoken. It has been suggested to us that instead of this binary classification of detected errors a more continuous variable such as time interval between onset and interruption might be more informative. However, in our corpus the distribution of detected errors as a function of the timing of interruption is highly bimodal. The amount of errors in between the CV…interruptions and the repaired fully spoken error forms is negligible (Nooteboom, 2005b).

Our starting materials for stimulus construction were the speech errors as produced by the speakers, with these errors being undetected or detected early or detected late, and with matching correct speech sounds differing from the misspoken segments. Hence, it turned out to be impossible to properly counterbalance consonant sounds (and hence, corresponding response keys) over detection categories. This may have allowed bias in our results: for example, if correct identification of a late-detected speech error would require pressing an infrequent or inaccessible key on the keyboard, then the responses may be biased against identification of the misspoken segment, not because of its late-detection status but perhaps in part due to the low usage frequency of the response key. Unfortunately our unbalanced distribution cannot neutralize this potential bias, but we note that the five most frequent target keys in our materials (*P*, *K*, *B*, *V*, *Z*), which together are involved in 486 out of 582 stimuli, have approximately the same key frequency in Dutch (Van den Broecke, 1988, p. 401). Hence the bias due to key frequency is likely to be small, if present at all. See also the analysis of reaction times in the Results section.

The reader may have noticed that our “controls” are not the same speech sounds as the corresponding misspoken segments. For instance the control segment of the speech segment *boo* from the speech error *boo…good beer* would have been *goo*. This is, of course, unfortunate, because we cannot rule out that the identifiability of the two segments may differ systematically. Given that we use a great many pairs of to-be-spoonerized consonants, in fact 42 different consonant pairs, we can only argue that differences in identifiability in specific pairs will average out over the whole set. However, as there was no phonemic identity of error segments and correct controls, we have to be cautious in drawing conclusions from comparisons involving the correct controls. Intensities of the speech fragments after these were excised from the original spoken word forms differed widely and disturbingly. Over the 140 speakers the way of speaking varied from whispering to nearly shouting. In fact, the experiment could hardly have been run successfully when keeping the original intensity differences intact. Therefore these intensities were normalized using Praat so that they were all in the order of 70 dB above threshold. It has been pointed out to us that in doing this, we potentially obliterated systematic differences in intensity between speech errors and correct controls that might be relevant to perception. However, as in this material there is no phonemic identity between speech errors and correct controls, there is no way to find out. It has also been pointed out to us that a major design limitation of this study is the lack of acoustic or articulatory measures of the stimulus materials. This is correct. However, pronunciations differed widely. Here also the lack of phonemic identity between error segments and their correct controls stood in the way of any attempt to use sensible acoustic measures to distinguish between segments that did and segments that did not suffer from articulatory blending.

Participants

Participants were 21 (18 females and 3 males) students and collaborators of Utrecht University, mostly from the Faculty of Humanities, ranging in age from 17 to 52 years. All participants were native speakers of Dutch, and none of them had a self-reported deficiency in speech, hearing, or in using a keyboard.

Procedure

Each participant, sitting in a sound treated booth, was presented over headphones first with 15 practice stimuli and then with all 582 speech fragments in the stimulus list in random order. Listeners' task was to identify the consonant sound and to react as fast as possible by typing the corresponding letter on a normal PC keyboard. When a letter had been typed on the keyboard, the following stimulus was presented after an interval of 1000 ms. On the keyboard all consonant letters except *Q* and *X* were enabled. *Q* and *X* and all vowel letters were disabled. The participants were informed about this and also knew that they had to press one of the enabled letters to continue the experiment. A PC monitor in front of the participant gave only a stimulus number, for the practice stimuli starting with 15 and counting down to 1 and for the test stimuli starting with 582 and counting down to 1. After the last stimulus had been responded to, a message appeared on the screen telling the subject that the experiment was over. The experiment lasted some 25 min and most participants found the task very easy. For each stimulus speech fragment the response was registered and the reaction time measured from the end of the speech fragment (always 40 ms after the vowel onset). Each participant was tested individually.[1]

Results

Error frequencies

As explained above, a misidentification is defined as an identification that deviates from the auditory transcription described in Nooteboom and Quené (2008). As can be seen in Table 2 below, although there are obviously more misidentifications in the speech error condition than in the correct controls, the great majority of all responses support the earlier auditory transcription. However, closer inspection of the actual misidentifications reveals that there is a potential problem with these data.

By far most misidentifications in both conditions appear to be misperceptions of voice, for example *p* heard as *b* or vice versa. Misperceptions of voice can only occur in our stimulus set with *b, p, d*, and *t*. Not with *k*, because *k* does not have a voiced counterpart in Dutch. And not with fricatives because for most students of the current generation that formed the majority of our participants the voice contrast has disappeared in initial fricatives (Van de Velde, Gerritsen, & Van Hout, 1995). Therefore confusions of voice in fricatives were not counted as errors, both in the speech error elicitation experiments and in the current identification experiment. Because we wished to include as many speech errors as possible in order to avoid a selection bias, our stimulus set was not controlled or balanced for voiced versus voiceless in initial stops. It turned out that in the condition *correct controls* there were 2752 responses and in the condition *speech errors* there were 3822 responses in which voice could be misperceived. This is the main reason why the number of misidentifications is so much greater in the condition *speech errors* than in the control condition. The abundance of voice errors in both conditions reflects the fact that in Dutch voice of stops in word initial condition is not perceptually very robust (Van Alphen, 2004). We have decided in our further data analysis not only not to consider misperceptions of voice in fricatives, as said above, but also not to consider mispercep-

[1] It should be noted that a keyboard is not an ideal instrument for measuring response or reaction times. One may find warnings (e.g. Segalowitz & Graves, 1990; Beringer, 1992) that, depending on the keyboard used, reaction times as measured with pressing the keys of a keyboard may include substantial additions to the participant's true reaction times and considerable increases in their variability. For this reason we have assessed the average delay and variability in measured time interval for the keys corresponding to the target consonants in the stimulus fragments, using a computer-controlled actuator that pressed each of these keys of the experiment keyboard 100 times (at 8 Hz). The average delay between the actuator movement and the computer-registered response was 15.5 ms (with $s = 3.4$ ms). Delays were significantly different between the target consonant keys [$F(9,990) = 3.63$, $p = .0002$]. Post-hoc inspection showed that this effect was due to the *P* key yielding shorter delays (−1.1 ms, *s.e.* 0.418) and the *S* key yielding longer delays (+1.5 ms, *s.e.* 0.418) than the other keys. However, these differences are extremely small [with partial $\eta^2 = 0.03$] and of the same magnitude as the 1-ms rounding error. For our purpose this average delay is not problematic, and, given the design of the experiment with many stimuli to be identified by many listeners, the extra variability in reaction times was judged to be acceptable.

Table 2
Comparison between speech errors and correctly spoken segments of the numbers of cases that consonantal fragments are misperceived. Percentages are given in brackets. 100% is the total nr of stimuli in that condition.

Stimulus	Nr. of misidentifications	Total nr. of stimuli
Correct controls	554 (9%)	6111
Speech error	952 (16%)	6111
Tot.	1506 (12%)	12,222

tions of voice in stops and to focus on other misidentifications instead. This implies that all confusions between voiced and voiceless counterparts were considered as correct identifications. Table 3 gives a breakdown of the remaining misidentifications. The responses were analyzed by means of mixed-effects logistic regression models (GLMMs), with misidentification as a binomial dependent variable (Quené & Van den Bergh, 2008). These models included two crossed random effects, viz. pairs of stimuli (speech error paired with matching control) and listeners (Baayen, Davidson, & Bates, 2008; Quené & Van den Bergh, 2008). The fixed effects of stimulus origin (control, undetected speech error, early detected speech error, late detected speech error) were analyzed as 3 planned orthogonal contrasts, corresponding to three predictions.

The first contrast compares the control stimuli against the speech errors. This contrast is computed by multiplying the log odds for the four respective stimulus origins, ordered as above, by the four respective contrast weights (−1, 1/3, 1/3, 1/3). This contrast therefore computes the difference between the (negative-signed) correct control condition, and the (positive-signed) average of the three (equally weighted) speech error conditions. The second contrast compares the undetected versus detected speech errors (contrast weights 0, −1, 1/2, 1/2); the third contrast compares the early versus late detected speech errors (weights 0, 0, −1, 1). In addition, the originally intended consonant phoneme was added as a fixed effect. GLMM analyses were performed using the packages lme4 (Bates, Maechler, & Bolker, 2012) and languageR (Baayen, 2012) for R (R Development Core Team, 2012). The results of the optimal GLMM are shown in Table 4. Interactions of origin by consonant were not included in the optimal model, because for most consonants there were too few misidentifications to investigate such interactions. Not surprisingly, the results of this GLMM show very low rates of misidentification. Responses which differ from the auditory transcription used in our earlier research (Nooteboom & Quené, 2008) are rare. Pooled over all stimuli, only 3% of the responses deviated from the earlier auditory transcription made by a single phonetically trained observer. In terms of identification, perceptual ambiguity is generally very low.

Table 4
Estimated coefficients of the optimal mixed-effects logistic regression model of misidentifications. Estimates of fixed coefficients are given in log odds, with standard errors and with significance levels. Estimates of random coefficients are given in standard deviations of log odds, with bootstrapped 95% confidence interval of the estimate. $N = 12{,}222$.

Fixed coefficients	Estimate	*s.e.*	*p*
Target.b	−3.9243	0.1972	.0001
Target.d	−3.2427	0.2909	.0001
Target.g	−4.8441	0.4677	.0001
Target.k	−4.9007	0.2527	.0001
Target.p	−5.4601	0.2635	.0001
Target.t	−4.8553	0.4810	.0001
Target.fv	−5.0548	0.2850	.0001
Target.z	−5.4134	0.3185	.0001
Control vs Speech error	+0.4655	0.1124	.0001
Undetected vs Detected	+0.4091	0.1553	.0085
Early vs Late Detected	+0.3068	0.1646	.0624
Random coefficients	Estimate		*N*
Listeners	0.4007	(0.4191, 0.8157)	21
Pairs (of stimuli)	1.6963	(2.1770, 4.0268)	291

Nevertheless, there are some perceptually ambiguous speech fragments, and the misidentification responses of these stimuli are distributed differently over the various stimulus categories. The three predicted effects of stimulus origin are indeed observed in the misidentification patterns, as shown in Table 4. The significant first contrast confirms that misidentification occurs more often in fragments derived from speech errors than in correct control fragments ($\beta = 0.4655$, odds ratio 1.59, $p = .0001$). The significant second contrast confirms that misidentification occurs more often in segmental errors detected by the speaker than in undetected errors ($\beta = 0.4091$, odds ratio 1.51, $p = .0085$). Although not highly significant, the third contrast suggests that late-detected errors are more often misidentified than early-detected errors ($\beta = 0.3068$, odds ratio 1.36, $p = .0624$).

From the articulatory research by Goldstein et al. (2007) and from the perceptual research by Pouplier and Goldstein (2005) one would expect that in speech fragments stemming from spoonerisms, the ambiguity is mainly between the error segment and the correct target. Of course,

Table 3
Number of misidentifications (other than those of voice) in the identification experiment, and number of responses where the *competing segment* was heard, i.e. the error segment in the *correct controls* condition and the correct target in the *speech errors* condition. Percentages are given between brackets. For the nr. of misidentifications 100% is the total nr of responses, for the *competing segments* 100% is the nr. of misidentifications. Agreement between participants is expressed as the average highest number of 21 participants who share the same perception over the speech fragments, divided by 21, times 100.

Stimulus origin	Nr. of misidentifications	Nr. of competing segment perceptions	Average agreement between participants (%)	Total nr. of responses
Correct controls	130 (2%)	6 (5%)	98	6111
Undetected errors	104 (3%)	13 (13%)	97	3318
Early detected errors	59 (4%)	6 (10%)	97	1680
Late detected errors	66 (6%)	7 (11%)	94	1113
Total	**359 (3%)**	**32 (8%)**	**97**	**12,222**

for the correctly spoken segments there is no such expectation. However, in Table 3 we have seen that responses involving the expected ambiguity with a specific, phonologically prepared, competing segment, are extremely rare. Perceptual ambiguity between a correct target and a phonologically prepared error segment, though apparently possible, is highly infrequent. It would be interesting to know whether there are significant differences in the numbers of *competing segment* perceptions between stimuli originating from undetected, early-detected and late-detected speech errors. Unfortunately these numbers are too low for further informative testing.

The main impression from these results is that misidentifications are rare, although they are more frequent for *speech errors* than for *correct controls*. In addition, *competing segment* perceptions, although obviously possible, are extremely rare. In terms of identification, perceptual ambiguity caused by the speech error generating process hardly ever occurs. Would it be possible that more traces can be found of perceptual ambiguity caused by speech errors, when we analyze reaction times measured in the identification task?

Identification times

Reaction times (RTs) were analyzed for a subset of the experimental stimuli. It turned out that many of the misidentifications in speech error stimuli (see Table 3) came from only a few of those stimuli. The distribution of misidentifications over speech error stimuli (Fig. 1) shows that 219 out of 291 stimuli were never misidentified, whereas there were also a few stimuli that were quite often misidentified. We suspect that those few items that yield many misidentifications are acoustically ambiguous; any differences in reaction times involving these speech error stimuli might then be attributed in part to these phonetically ambiguous speech signals yielding many misidentifications. In order to exclude this possible explanation of a possible RT difference, speech error stimuli yielding more than 3 misidentifications (across 21 listeners) were excluded from the RT analysis. This criterion excluded 19 speech error stimuli; their matching correct controls were also excluded. The remaining 11,424 responses contained 198 misidentifications (2%); these were also excluded from further analyses. Hence the RTs analyzed below were collected on correct responses only, and on stimuli that were seldom misidentified.

The reaction times of the remaining responses were log-transformed to obtain a more normal distribution. Log-transformed outlier RTs were cut off at a value 3 times the interquartile range above the third quartile (Tukey, 1977); this cutoff value corresponds with a raw RT of 2578 ms. Using this cutoff value, 29 outlier RT observations were discarded. These outliers were distributed equally over listeners (10 listeners with no outliers, 3 listeners with 1 outlier, 2 listeners with 2 outliers, 4 listeners with 3 outliers, and 2 listeners with 5 outliers) and over stimulus pairs (245 pairs with no outliers, 25 pairs with 1 outlier, and 2 pairs with 2 outliers).

The log-transformed RTs were analyzed by means of a mixed-effects regression model (LMM), with two crossed random effects, viz. pairs of stimuli (speech error paired with matching control) and listeners (Baayen et al., 2008; Quené & Van den Bergh, 2008). As in the previous analysis, the fixed effects of stimulus origin (control, undetected speech error, early detected speech error, late detected speech error) were analyzed as three planned orthogonal contrasts corresponding to three predictions. The first contrast compares the control stimuli against the speech errors (weights -1, 1/3, 1/3, 1/3). The second contrast compares the undetected versus detected speech errors (weights 0, -1, 1/2, 1/2); the third contrast compares the early versus late detected speech errors (weights 0, 0, -1, 1). The originally intended consonant phoneme and the trial number (centered to its median) were also included as fixed effects.

The random part of the LMM also contained the stimulus origin, so that homogeneity of between-stimulus variance was not assumed (Quené & Van den Bergh, 2004), and it contained the trial number effect, so that the listeners' training slope (i.e. linear effect of trial number on RT) was allowed to vary between listeners. The inclusion of these terms into the random part of the model was warranted by Likelihood Ratio Tests. The optimal model reported in Table 5 was also compared against a more complex model, containing interactions of stimulus origin and intended phoneme. However, the more complex model did not perform better, according to a Likelihood Ratio

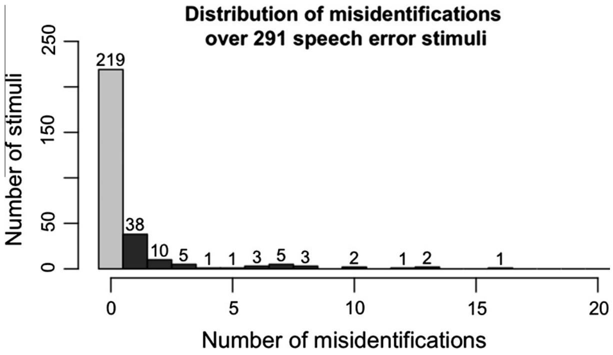

Fig. 1. Distribution of misidentification responses by 21 listeners, over 291 speech error stimuli.

Table 5
Estimated coefficients of the optimal mixed-effects regression model of log-transformed response times. Estimates of fixed coefficients are given in log ms units, with standard errors and with significance level. Estimates of random coefficients are given as variance–covariance matrices. $N = 11{,}197$.

Fixed coefficients	Estimate	*s.e.*	*p*
Target.b	6.5379	0.0279	.0001
Target.d	6.5826	0.0318	.0001
Target.g	6.4075	0.0338	.0001
Target.k	6.4081	0.0283	.0001
Target.p	6.5801	0.0279	.0001
Target.t	6.5893	0.0340	.0001
Target.fv	6.5783	0.0286	.0001
Target.z	6.6595	0.0293	.0001
Control vs Speech error	−0.0027	0.0064	n.s.
Undetected vs Detected	−0.0023	0.0074	n.s.
Early vs Late Detected	+0.0218	0.0088	.0191*
TrialNr (Centered).	−0.000056	0.000055	n.s.

	Random between pairs (of stimuli) (*N* = 272)			
	.Ctrl	.Undet	.Early	.Late
.Ctrl	0.00398			
.Undet	0.00001	0.00308		
.Early	0.00077	0.00000	0.00265	
.Late	0.00067	0.00000	0.00013	0.00671
Random between listeners (N = 21)				
	intercept	TrialNr		
intercept	0.01492			
TrialNr	0.00001	0.00000		
Residual (N = 11197)				
	0.52713			

Test, than the optimal model summarized in Table 5. The resulting regression coefficients in Table 5 show a significant main effect of the origin of the speech fragment. The first and the second contrast contained in this main effect are not significant, the third contrast, between early and late detected is significant. In order to exclude the key frequency of the response keys as a possible confound, the same analysis was also run on a subset of the materials, involving only the keys P, K, B, V, and Z, which were the most frequent response keys in our materials, and which have a similar average key frequency in Dutch (see Materials subsection). This LMM analysis yielded a pattern very similar to the one for the full set of stimuli (as summarized in Table 5): the first and second contrasts were again not significant, and the third contrast, between early and late detected errors, did again yield a significant effect ($\beta = +0.0284$, *s.e.* 0.0093, $p = .0057$).

The first contrast compares reaction times to correct control fragments with those to fragments excised from speech errors. The average RTs in Fig. 2 suggest that this contrast is insignificant because of the considerable differences in reaction times between fragments taken from undetected, early detected and late detected speech errors, and particularly because fragments taken from early detected speech errors yield marginally shorter, and not longer reaction times than fragments taken from correct controls ($\beta = -0.0266$, $p = .0182$, according to a post hoc comparison of these conditions). Similarly, the second contrast, between undetected and detected speech errors, yields an insignificant difference in reaction times because

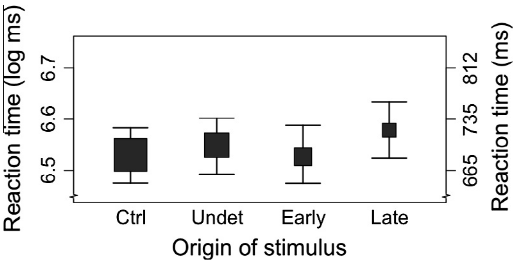

Fig. 2. Average log-transformed reaction times, broken down by stimulus origin (control, undetected, early detected, late detected). Error bars correspond to 95% confidence intervals of the corresponding bootstrapped average reaction times, over 250 two-stage bootstrap iterations. Symbol sizes correspond to the number of responses per cell. Note that the observed average reaction times may deviate from the center of the bootstrap confidence intervals.

early detected speech errors lead to shorter reaction times than undetected errors ($\beta = -0.0252$, $p = .0295$, according to a post hoc comparison) whereas late detected speech errors lead, with marginal significance, to longer reaction times than undetected errors ($\beta = +0.0184$, $p = .0734$, according to a post hoc comparison). The significant third contrast indicates that fragments from errors detected late by the speaker require significantly longer RTs than fragments from errors detected early ($p = .0191$, see Table 5). After back-transformation, this contrast corresponds with a difference of about 721–685 = 36 ms in average reaction time.

These findings are surprising. The reaction times suggest that, as expected, speech fragments stemming from *early detected* segmental errors are on average perceptually clearer, less ambiguous, than those stemming from *undetected errors*. In addition the reaction times to *early detected* errors tend to be, contrary to expectation, not longer but shorter than reaction times to *correct controls* (but the reader will recall that we should be cautious in drawing conclusions from comparisons involving the correct controls, because of a lack of phonemic identity between speech errors and controls). On the other hand speech fragments originating from *late detected* segmental errors are not only, as expected, perceptually less clear, or more ambiguous, than *correct controls* and *early detected* errors, but also, contrary to expectation are perceptually less clear than *undetected errors*. We will come back to these unexpected findings below in the discussion of the results.

Discussion

The current experiment was set up to find out whether or not, in a speech fragment identification task, systematic differences could be found between speech fragments excised from correct controls and speech fragments excised from speech errors, and between speech fragments excised from not detected and repaired, early detected and repaired and late detected and repaired speech errors. That there might be such differences between speech sounds was predicted on the basis of results by Goldstein et al. (2007) and McMillan and Corley (2010). These authors found in articulatory measurements on segmental speech errors elicited with metronome-controlled tongue twisters and their correct controls that most elicited speech errors were not categorical, but rather gradient with simultaneous articulatory gestures from competing phonemes. Pouplier and Goldstein (2005) demonstrated in a phoneme decision experiment with stimuli stemming from those articulatory experiments that competing phonemes can be less well discriminated and lead to longer reaction times in segments containing conflicting articulatory gestures than in correct controls. In the current identification experiment we used speech fragments excised from segmental speech errors and their correctly produced controls, as elicited with the SLIP technique in two experiments described by Nooteboom and Quené (2008). We assumed that speech errors providing perceptually clearer, less ambiguous, speech fragments are easier to detect in self-monitoring for segmental errors than speech fragments from perceptually less clear, more ambiguous speech errors, and we made a number of straightforward predictions to be discussed below. An unexpected finding was that fragments excised from early detected errors have reaction times that are, if anything, not longer but shorter than those from correct controls, therefore suggesting that these speech sounds do not carry traces of articulatory blending. Another unexpected finding was that late detected errors not only have, as predicted, longer reaction times than fragments excised from early detected errors, but that these fragments from late detected error segments have the longest reaction times of all conditions, even longer than those of fragments from undetected errors. Below we will discuss each of the predictions we made and the corresponding results.

- (1) Speech fragments excised from segmental speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from correct controls.

A first obvious result concerning misidentifications is that they are rare events. We find misidentifications in only 3% of the responses (but note that this excludes all misidentifications of voice). The overwhelming majority of responses supports the auditory transcription made by a single phonetically trained observer as reported in Nooteboom and Quené (2008). This is an important result because it means that auditory transcriptions both of segmental errors in spontaneous speech and of experimentally elicited speech errors, as used by many researchers ever since the pioneering work by Meringer and Mayer (1895) and Meringer (1908), can be basically reliable as descriptions of what people actually perceive when segmental speech errors are made. It should be noted, though, that this finding does not tell us much about the probability that either in the speech errors or in the correct controls, pronunciation is affected by articulatory blending. In this respect the analysis of reaction times may be more relevant. But despite the low frequency of misidentifications in our experiment, the prediction that such misidentifications are more frequent for fragments excised from speech errors than for fragments excised from correct controls is borne out. This suggests that speech fragments taken from segmental errors are more often perceptually unclear than those taken from correct controls. However, because the set of error segments and the set of control segments consisted of different phonemes, it cannot be excluded that this result is an artifact of the phoneme sets being compared. It should also be noted that the cases concerned form only a tiny fraction of all error-derived speech fragments. In the overwhelming majority of cases, all listeners confirmed the auditory transcription described in Nooteboom and Quené (2008).

If many segmental speech errors do indeed carry traces of simultaneous competing articulatory gestures, whereas there is no reason to suppose the same for correct controls, then one expects that speech sounds resulting from segmental errors on average are perceptually less clear than speech sounds resulting from their correct controls. Reaction times increase as perceptual ambiguity increases (Botvinick et al., 2001; Szmalec et al., 2008), and therefore we would expect that speech fragments excised from speech errors yield longer reaction times than those excised from correct controls. However, we find that there is no significant difference in reaction times between fragments from correct controls and fragments from speech errors. Reaction times for undetected errors and late detected errors are, as predicted, longer than those for correct controls, but the predicted effect was nevertheless absent because reaction times for early detected errors are not longer but even shorter than those for correct controls. This suggests that on average fragments from early detected errors are perceptually at least just as clear as fragments from correct controls. If valid, this would be an important finding. It would suggest that these early interrupted speech errors do not suffer at all from articulatory blending. This also would mean that segmental speech errors may begin their life in inner speech as full categorical substitutions. We will attempt to explain the special status of early detected errors when discussing the difference between early and late detected errors.

- (2) Speech fragments excised from undetected speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from detected speech errors.

As predicted undetected errors suffer significantly more from misidentification than detected errors. This seems to suggests that indeed, as we had assumed, detected speech errors in more cases have led to perceptually clear versions of the erroneous segments than undetected speech errors did. We had predicted this under the assumption that in self-monitoring comparison between target and error word form is the major source of detection of segmental errors. This assumption seems supported by the current finding. But it should be kept in mind that misidentifications are so infrequent that their distribution does not necessarily tell us much about the great bulk of error segments. Reaction times in correct identifications, the bulk of responses, tell a different story.

If indeed, on average, speech fragments from undetected speech errors were perceptually less clear than those excised from detected errors, then one would expect that reaction times to fragments excised from undetected errors are significantly longer than those to fragments taken from detected errors. This is not the case. As we have seen in the results section, reaction times to fragments excised from early detected errors are shorter and reaction times to fragments excised from late detected errors are longer than those to undetected errors. There appears to be a major difference between early and late detected errors that can best be discussed in the context of our third prediction.

- (3) Speech fragments excised from late detected speech errors will be more often misidentified and will on average have longer reaction times than speech fragments excised from early detected speech errors.

This prediction was derived from a perception-based monitor for speech errors, and particularly from the proposal by Nooteboom and Quené (2008) that monitoring for speech errors employs as its main tool perceptual comparison between error form and target form. This would suggest that the odds of error detection increases with perceptual distance between error and target, and that early error detection precisely is early because the perceptual distance between error and target is relatively great, whereas late error detection precisely is late because the perceptual distance between error and target is somewhat less. Therefore we predicted that late detected errors are more often misidentified and have longer reaction times than early detected errors. We have seen that late detected errors indeed lead to more misidentifications than early detected errors, although the difference is only marginally significant. Misidentifications are so rare that the differences found may not be representative for the effects of relative perceptual clarity on the great bulk of speech errors and correct controls. In that respect the reaction times in correct identifications may provide more insight.

Our results show that indeed fragments excised from late detected errors lead to significantly longer reaction times in identification than fragments excised from early detected errors. However, the pattern of the reaction times is very different from what we had expected. We had predicted that both early and late detected errors would lead to shorter reaction times than undetected errors. This would have supported the proposal by Nooteboom and Quené (2008) that monitoring for speech errors depends on perceptual comparison between error form and target form. In contrast, however, early detected errors lead to shorter reaction times than undetected errors and late detected errors lead to longer reaction times than undetected errors. There appears to be a major difference between early and late detected errors in perceptual clarity. Early detected errors behave as if they do not suffer at all from articulatory blending, whereas late detected errors behave as if they suffer more than any other category of errors from articulatory blending. This unexpected finding begs to be explained.

The pattern of reaction times seems to suggest that early detected errors betray the effect of perception-based monitoring and late detected errors betray the effect of conflict-based monitoring. This does not seem very parsimonious. Let us begin with the early detected errors or early interruptions of the type *boo*…*good beer*. Nooteboom and Quené (2008) suggested that such early interruptions result from too hasty speech initiation. Speech would be initiated before self-monitoring could have detected the error. But very shortly after speech is initiated self-monitoring catches up, the error is as yet detected and speech is interrupted. The fact that early interruptions do not seem to suffer at all from articulatory blending now suggests to us that in these cases speech is initiated before the correct target segment is inserted into the speech plan as a competitor of the error segment. This seems entirely possible if we assume that, when two segments are activated for the same slot in the speech plan, the segment with the highest activation is inserted slightly earlier than its competitor. The class of early interruptions would then consist of those cases where the error segment is activated somewhat more strongly than the competing correct target segment. As soon as the error segment is inserted, at time *t*, speech is initiated. Immediately thereafter, at time $t + 1$, the correct target segment also reaches the speech plan and there arises a conflict between the two rivaling segments, competing for the same slot. Because of this conflict the error is detected and speech is interrupted. This speculative view of the origin of early interrupted speech, wholly in line with the model of conflict monitoring proposed by Botvinick et al. (2001) and Yeung et al. (2004), explains why these speech errors do not suffer from articulatory blending. The articulatory blending is absent not because there is no conflict between the two segments, but because the conflict only arises after speech is already initiated. This may reconcile the absence of articulatory blending in these cases with a conflict-based theory of monitoring for speech errors. It seems likely that in the experiments by Goldstein et al. (2007) and McMillan and Corley (2010) this class of early interrupted speech errors, not suffering from articulatory blending, is absent, simply because under the conditions of these experiments people seem to make no or very few self-repairs, neither early nor late. Crucially our findings show that a segmental speech error may begin its life in inner speech as a full substitution, not yet affected by conflict with its upcoming rival.

Fragments from late detected errors on average lead to longer reaction times than fragments from any other category of errors, and therefore must on average have been perceptually less clear than all other errors. This suggests that the amount of conflict between competing segments on average is greater for late detected than for undetected errors. This result seems to support some form of conflict-based monitoring as proposed by Nozari et al. (2011). The reader may note, however, that, where the conflict-based theory of speech monitoring is about speech preparation, late detected errors very likely are not detected during speech preparation. They may be detected on the basis of efferent commands, tactile information or proprioception of articulatory gestures during articulation, employing as criterion a measure of conflict between competing articulatory gestures (cf. Postma, 2000). Another possibility is that these errors are detected in overt speech (cf. Levelt, Roelofs, & Meyer, 1999; Hartsuiker et al., 2005; Huettig & Hartsuiker, 2010; Nooteboom, 2005a, 2010). In that case perceptual monitoring may react to the perceptual unclarity of the speech sounds concerned. It should be noted that this would constitute some form of perception-based monitoring, although not on the basis of comparison between error and target form. This would be a form of perception-based conflict monitoring. A possible way to distinguish between the two interpretations would be to repeat the SLIP experiments of Nooteboom and Quené (2008), but this time with perception of overt speech made inaccessible to the speaker by auditory masking with loud noise. If the pattern of late detected errors would be the same as in Nooteboom and Quené (2008), this would plead for production-based monitoring of articulation. If late detected errors would become very rare or disappear in the noise condition, this would suggest that late detected errors are detected in overt speech. It is noteworthy that several studies using the SLIP technique or an equivalent method had their speakers listen to loud noise: Baars et al. (1975), Hartsuiker, Corley, and Martensen (2005), McMillan (2008). However, it is not clear what the effect of this noise on the behavior of the speakers was. Also, none of these studies had a breakdown of the results in terms of early and late detected errors. The only two studies with normal speakers we are aware of in which the absence or presence of noise was an experimental variable are reported by Lackner and Tuller (1979) and by Postma and Noordanus (1996). Lackner and Tuller found that the number of self-repairs significantly decreased under auditory masking as compared to a condition without auditory masking. This pleads for a contribution of perception-based error detection in overt speech. However, these authors did not distinguish between early and late detected errors. It would still be possible that speech errors are detected both during articulation and in overt speech. Postma and Noordanus, who asked their subjects to report their errors made during the speeded production of tongue twisters, found that more errors are reported by their subjects with than without auditory feedback. However, they did not look into self-repairs. All in all, although the current results currently do not seem to give firm ground for a choice between perception-based and production-based monitoring, to us it seems most parsimonious to assume conflict-based monitoring both for inner speech and for articulation or overt speech.

In summary, properties of correct and speech error segments stemming from SLIP experiments eliciting segmental speech errors can be meaningfully studied by having listeners identify speech fragments excised from the correct forms and error forms. We find that misidentifications, defined as deviations from the auditory transcription by a single trained phonetician, are very rare, in the order of 3%, but that these rare misidentifications are on average more frequent for fragments derived from speech errors than for fragments derived from correct controls. We also find that misidentifications of fragments taken from undetected errors are more frequent than those of fragments taken from detected errors, and more frequent for fragments from late detected errors than for fragments from early detected errors. We conclude from these findings on error frequencies that auditory transcription of experimentally elicited speech errors is reasonably reliable as a description of what people perceive, but also that perceptual unclarity caused by articulatory blending is possible. The data on misidentifications tell us very little about the frequency of articulatory blending in speech errors. For speech fragments excised from speech errors but not showing any sign of perceptual ambiguity in their misidentification scores, reaction times betray systematic differences in perceptual clarity between fragments taken from correct controls and those taken from undetected errors and late detected errors. The most salient findings are that, contrary to prediction, early detected errors are at least as perceptually clear as correctly spoken segments, and that, also contrary to expectation, late detected errors are perceptually less clear than all other classes of stimuli. The perceptual clarity of early detected errors suggests that (a) segmental errors originate in inner speech as full substitutions, and (b) in these early detected errors speech was initiated before the competing correct targets were activated. After activation of the correct target the error may be detected on the basis of conflict between the two competing segments, and speech will then be interrupted early. Late detection either reflects self-monitoring of overt speech, reacting to the perceived unclarity of a speech sound, or self-monitoring of articulation, possibly employing efferent commands, tactile information and proprioception of articulatory gestures, and using the presence of simultaneous competing articulatory gestures as a criterion for error detection. The latter interpretation would be in accordance with some form of conflict monitoring (Nozari et al., 2011). For now, we conclude from our identification experiment that early and late detection in self-monitoring reflect monitoring of different stages of speech production, possibly employing different criteria in error detection.

Acknowledgments

Our thanks are due to Theo Veenker for technical assistance and for measuring the delays and variability in reaction times for the different keyboard keys used in the experiment. The raw data of the experiment are currently available on-line in the form of an excel document at the following web adress: http://www.let.uu.nl/~Sieb.Nooteboom/personal/Experimentaldata.htm.

It should be noted that this Excel file also provides the stimulus properties that are relevant to the analysis presented in this paper.

References

Baars, B. J., & Motley, M. T. (1974). Spoonerisms: Experimental elicitation of human speech errors. *Journal Supplement Abstract Service, Fall 1974. Catalog of Selected Documents in Psychology, 3*, 28–47.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14*, 382–391.

Baayen, R. H. (2012). *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. R package version 1.4. <http://CRAN.R-project.org/package=languageR>.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0. <http://CRAN.R-project.org/package=lme4>.

Beringer, J. (1992). Timing accuracy of mouse response registration on the IBM microcomputer family. *Behavior, Research Methods, Instruments, and Computers, 24*, 486–490.

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. (2009). *Praat: Doing phonetics by computer (Version 5.1.05)* [Computer program]. <http://www.praat.org/>.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychology Review, 108*, 624–652.

Butterworth, B., & Howard, D. (1987). Paragrammatism. *Paragrammatism, 26*, 1–37.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30*, 139–162.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649–683.

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

Hartsuiker, R., 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). *Journal of Memory and Language, 52*, 58–70.

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

Hartsuiker, R. J., Kolk, H. H. J., & Martensen, H. (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* (pp. 187–205). Hove: Psychology Press.

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes, 25*, 347–374.

Lackner, J. R., & Tuller, B. H. (1979). Role of efferent monitoring in the detection of self-produced speech errors. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 281–294). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Levelt, W. J. M. (1989). *Speaking. From intention to articulation*. Cambridge, Massachusetts: The MIT Press.

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

Liss, J. M. (1998). Error-revision in the spontaneous speech of apraxic speakers. *Brain and Language, 62*, 342–360.

Marshall, J., Rapaport, B. Z., & Garcia-Bunuel, L. (1985). Self-monitoring behavior in case of severe auditory agnosia with aphasia. *Brain and Language, 24*, 297–313.

Marshall, J., Robson, J., Pring, T., & Chiat (1998). Why does monitoring fail in jargon aphasia? Comprehension, judgment, and therapy evidence. *Brain and Language, 63*, 79–107.

McMillan, C. T. (2008). *Articulatory evidence for interactivity in speech production*. Unpublished doctor's thesis. University of Edinburgh.

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

Meringer, R. (1908). *Aus dem Leben der Sprache*. Berlin: V. Behr's Verlag.

Meringer, R., & Mayer, K. (1895). *Versprechen und Verlesen*. Stuttgart: Goschensche.

Mowrey, R., & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*, 1299–1312.

Nooteboom, S. G. (2005b). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication, 47*, 43–58.

Nooteboom, S. G. (2010). Monitoring for speech errors has different functions in inner and overt speech. In M. Everaert, T. Lentz, H. De Mulder, A. Nilsen, & A. Zondervan (Eds.), *The linguistic enterprise: From knowledge of language to knowledge in linguistics* (pp. 213–235). Amsterdam: John Benjamins Publishing Company.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand* (pp. 87–95). New York: Academic Press.

Nooteboom, S. G. (2005a). Listening to one-self: Monitoring speech production. In R. Hartsuiker, Y. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 167–186). Hove: Psychology Press.

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. *Journal of Memory and Language, 58*, 837–861.

Nozari, N., Dell, G., & Schwartz, M. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63*, 1–33.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97–131.

Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*(4), 375–392.

Pouplier, M. (2007). Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech, 50*, 311–341.

Pouplier, M., & Goldstein, L. (2005). Asymmetries in the perception of speech production errors. *Journal of Phonetics, 33*, 47–75.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication, 43*, 103–121.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413–425.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. <http://www.R-project.org/>.

Segalowitz, S. J., & Graves, R. (1990). Suitability of the IBM XT, AT and PS/2 keyboard, mouse, and game port as response devices in reaction time paradigms. *Behavior Research Methods, Instruments, and Computers, 22*, 283–289.

Szmalec, A., Verbruggen, F., Vandierendonck, A., De Baene, W., Verguts, T., & Notebaert, W. (2008). *Neuroscience Letters, 435*, 158–162.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Alphen, P. M. (2004). *Perceptual relevance of prevoicing in Dutch*. Unpublished doctoral dissertation. Nijmegen, the Netherlands: Radboud University.

Van de Velde, H., Gerritsen, M., & Van Hout, R. (1995). De verstemlozing van de fricatieven in het Standaard-Nederlands., Een onderzoek naar taalverandering in de periode 1935–1993 [The devoicing of fricatives in Standard Dutch. An investigation of language change in the period 1935–1993]. *De Nieuwe Taalgids, 88*, 422–445.

Van den Broecke, M. P. R. (1988). Frequenties van letters, lettergrepen, woorden en fonemen in het Nederlands [Frequencies of characters, syllables, words and phonemes in Dutch]. In M. P. R. van den Broecke (Ed.), *Ter Sprake: Spraak als betekenisvol geluid in 36 thematische hoofdstukken* (pp. 400–415). Dordrecht: Foris.

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