Do speakers try to distract attention from their speech errors? The prosody of self-repairs.

Sieb Nooteboom and Hugo Quené
Utrecht institute for Linguistics OTS, Utrecht University

Do speakers try to distract attention from their speech errors? The prosody of self-repairs.

Sieb Nooteboom and Hugo Quené
Utrecht institute for Linguistics OTS, Utrecht University

Abstract

Self-repairs of segmental speech errors come in two varieties: repairs of early and of late-detected errors. Early-detected errors are those where the error word is interrupted immediately after the error segment, late-detected errors are those in which the word or phrase containing the error is completed before the repair is made. We measured duration, maximum pitch, average pitch, maximum intensity, average intensity and spectral slope of the first vowel of both the reparandum (the stretch of speech containing an error and to be replaced by the repair, cf. Levelt 1983) and the repair, and measured offset-to-repair times, for repairs of early- and late-detected errors separately. Main findings were that repairs of early detected errors had higher average and maximum intensity, less steep spectral slope and much shorter offset-to-repair times than repairs of late-detected errors. We also ran a listening experiment on pairs of CV fragments excised from reparandum and repair, in which listeners were asked to indicate which of the two fragments was loudest. CV-fragments from repairs of early-detected errors were on average judged to be louder than the reparandum fragments. In repairs of late-detected errors, however, the subjective difference tends to be reversed. These findings suggest that in repairs of early-detected errors speakers hasten to shift the listener's attention from the error to the repair, whereas in repairs of late-detected errors speakers do not attempt to shift the listener's attention but rather provide a de-emphasized revision of their erroneous speech.

1. Introduction

In spontaneous speech roughly half of all segmental errors of speech (e.g. ***good beer >> bood beer***) are detected and repaired by the speaker (Nooteboom, 1980; Nooteboom, 2005a). Also experimentally elicited segmental errors of speech are often repaired by the speaker, although the repair frequency depends on the detailed structure of the task (Nooteboom and Quené, 2008; Nooteboom and Quené, 2013a). Repairs come in at least two varieties: Repairs of early-detected errors, i.e. repairs following speech errors that are immediately followed by speech interruption, often in midword (e.g. ***boo..good beer)***, and repairs of late-detected errors, i.e. repairs following speech errors in which the erroneous word or phrase is fully realized (e.g. ***bood beer..uh..good beer***). Obviously, in early-detected errors the speaker takes much less time after the error to interrupt speech for making a repair than in late-detected errors (Nooteboom, 2005b). One might have expected that this difference in waiting time is compensated for in the offset-to-repair time, so that the total time interval between error and repair would be roughly the same. However, this is not the case. The offset-to-repair times are considerably and significantly shorter in early-detected errors than in late-detected errors, often being in the order of 0 ms (Blackmer and Mitton, 1991; Nooteboom, 2005b). The rapid interruptions demonstrate that early-detected errors are not detected in overt speech but must have been detected in inner speech, because the erroneous speech fragments are often shorter than humanly possible reaction times. The very brief offset-to-repair times, too short to be used for planning a repair, suggest that interruption does not immediately follow error detection but rather is postponed until a repair of the internally detected error has been planned (Seyfeddinipur, Kita and Indefrey, 2008). For late-detected errors the relatively long error-to-offset time and the relatively long offset-to-repair time suggest that in these cases most likely errors are only detected after speech was initiated and that the repair generally is planned only after speech has stopped.

In a study of speech errors in spontaneous speech that were acoustically recorded, Cutler (1983) confirmed an observation by Goffman (1981) and reported that self-repairs come in two classes, those that are and those that are not prosodically marked. She also observed that "marking" is applied to errors at the lexical level or above. Repairs of phonetic errors were never prosodically "marked". Levelt and Cutler (1983), on the basis of auditory analysis of speech errors made in an interactive network description task and acoustically recorded, reported that "marked" and "unmarked" repairs of interrupted errors were roughly equally frequent, whereas "unmarked" repairs were slightly more frequent for completed errors. "Marked" errors are much more frequent for error repairs than for appropriateness repairs. No distinction was made between sound level errors and word level errors. Shattuck-Hufnagel and Cutler (1999), on the basis of an acoustic and auditory investigation of some 90 repaired speech errors of which they had acoustic recordings, reported that prosodic "marking" is more likely for word-level errors than for sound-level errors. Nooteboom (2010) measured maximum pitch and maximum intensity in the first vowel of segmental speech errors and their repairs, separately for repairs of early- and late-detected errors. He found that on average repairs of early-detected errors have higher pitch and greater intensity than repairs of late-detected errors. Note that this finding does not seem to tally with Cutler's observation that sound-level errors are never prosodically "marked". Nooteboom (2010) interpreted his finding as showing that speech prosody is used by speakers to distract the listener's attention from the early detected error and to attract attention to the late-detected error and its repair. Levelt and Cutler (1983) rather speak of "prosodic marking" as a signal that the error is "rejected". They assume that, where there is no "prosodic marking" of the repair, very likely the prosody of the repair is a close replica of the prosody of the reparandum and this would help the listener to identify the reparandum and replace it with the repair.

Plug (2011) investigated phonetic aspects of self-repairs of speech errors in Dutch, measuring speech rate and spectral reduction in reparandums and repairs of speech errors found in a corpus of recorded unscripted speech in Dutch (Ernestus, 2000). His main finding was that there was no difference between error repairs and appropriateness repairs, seemingly contradicting Levelt and Cutler (1983). However, it should be noted that Plug, in not measuring pitch and/or intensity, did not investigate prosodic marking. Neither did he distinguish between repairs of early- and late-detected errors, which were found to have different properties both in spontaneous speech and in speech error elicitation experiments by Nooteboom (2010).

The current investigation can be seen as an extension of Nooteboom (2010), limited to experimentally elicited segmental speech errors. The self-repairs reported on below stem from a new speech error elicitation experiment. In addition to maximum pitch and maximum intensity, acoustic measures used to compare reparandum and repair include average pitch, average intensity, duration, spectral slope and offset-to-repair times; These measures were compared separately for early-detected and late-detected segmental errors. In order to find out whether the differences found could have been used as signals in speech communication we ran a listening experiment with naive listeners comparing the perceived loudness of the first CV speech fragment of reparandum and repair, separately for early- and late-detected errors.

2. Materials

The materials for this investigation stem from a somewhat unusual SLIP experiment (Baars, Motley and MacKay, 1975). Generally, in a SLIP experiment exchanges are elicited between two word initial consonants, as in **good beer** turning into **bood geer**. Exceptions are Humphreys (2002), who elicited both exchanges and anticipations of initial consonants of English CVC words, and Nooteboom and Quené (2013a) who elicited exchanges, anticipations and perseverations of initial consonants, of Dutch CVC words. In the (as yet unpublished) experiment from which materials were taken for the current investigation we elicited exchanges, anticipations, and perseverations of C_1, V and C_2 in pairs of Dutch CVC words. The goal of the experiment was to study the assumed predominance of C_1-speech errors (Shattuck-Hufnagel, 1987). Here we will not discuss this any further. We will focus on those segmental substitutions that were clearly repaired. Repairs of late-detected errors consisted of repairs where the reparandum was completed. This gave us 57 repairs of early-detected errors and 37 repairs of late-detected errors. Table 1 shows how these repaired speech errors were distributed over the three positions in which speech errors were elicited.

Table 1. Distribution of repairs of early- and late-detected errors over three positions in CVC words in which segmental speech errors were elicited.

	C_1	V	C_2	Total
early	25	16	16	57
late	10	10	17	37
total	36	26	33	94

Table 2. Distribution of repairs of early and late-detected over three error types that were elicited.

	EXCH	ANTIC	PERSEV	Total
early	29	21	7	57
late	21	8	8	37
total	50	29	15	94

Table 2 shows how these repaired speech errors were distributed over the three types of speech errors that we attempted to elicit. It should be noted that these numbers do not reflect how many exchanges, anticipations and perseverations were actually made. This we do not know, because an early-detected error is defined as an early interruption, and it is unknown whether an early interruption such as **boo..good beer** stems from an exchange or an anticipation in inner speech. Also we can not know whether repaired perseverations of the C_2, like **pack back..pack bat** are early or late repairs. Table 3 shows the distribution of responses over exchanges, anticipations, perseverations and early interruptions that were actually made.

Table 3. Numbers of segmental substitutions for different error types.

	exchanges	anticipations	perseverations	other	early interruptions	Total
error type	25	6	1	15	47	94

In table 3 we see that in the actual speech errors made exchanges far outnumber anticipations and perseverations. It has been argued that the predominance of exchanges over other segmental substitutions made in inner speech is a highly robust phenomenon, and that most early interruptions stem from half-way repaired exchanges (Nooteboom and Quené, 2013a). The set of repaired segmental speech errors described here forms the material for our further investigation of the prosody of self-repairs.

3. Acoustic study

3.1. Measurements

The set of repaired speech errors described above, contains speech errors in all three positions of CVC words. For the errors in V position this implies that reparandum and repair had different vowels. This is unfortunate for a comparison of acoustic properties, because most acoustic properties we intend to compare, viz. duration, maximum pitch, average pitch, maximum intensity, average intensity and spectral slope of the first vowel of both the reparandum and the repair, do inherently depend on vowel identity. For comparing reparandum and repair with respect to these acoustic properties, we therefore made a further selection of repaired segmental errors, by excluding all vowel errors. We also excluded all cases in which the repair was interrupted before the first vowel was completed. This gave for the analysis the numbers as shown in Table 4.

Table 4. Distribution of early and late repairs as used in the acoustic analysis over C_1 and C_2 .

	C_1	C_2	Total
early repairs	22	16	38
late repairs	10	15	25
total	32	31	63

Using PRAAT (Boersma and Weenink, 2009), maximum pitch, average pitch, maximum intensity, average intensity and spectral slope (cf. Van Heuven and Sluijter, 1996; Sluijter and Van Heuven, 1996) of the initial vowels in both reparandum and repair were measured. Here spectral slope is defined as the power in a log-weighted combination of bands 1600−3200 and 3200−4800 Hz, minus the power in band 0−1600 Hz, expressed in dB. We also measured offset-to-repair times. These data were fed into several analyses. In the first set of analyses, dependent variables consisted of the various phonetic properties of the vowel in the *repair*. Separate linear mixed effects models (LMMs) were estimated for each of these dependent variables (Baayen, Davidson and Bates, 2008; Quené and Van den Bergh, 2004, 2008), with speakers (n=38) as a random effect[1]. LMM estimates were obtained using the package *lme4* (Bates, Maechler and Bolker, 2013) in *R* (R Core Team, 2013). Predictors in the LMMs were the same phonetic property of the corresponding vowel in the *reparandum*, as well as the detection status of the speaker's mispronunciation (dummy coding, code 0=late, n=25 cases, code 1=early, n=38 cases), and the time interval between the offset of the reparandum and the onset of the repair (offset-to-repair time), log-transformed and centered to its median. The interaction between these two predictors was also included as a predictor. The significance of the detection status predictor was assessed by means of t tests using d.f. procedure to estimate degrees of freedom (here 38−4−1 d.f.). We also report the 95% confidence interval of this detection status effect; this summarizes the estimated difference (in ms, in Hz, or in dB) between repairs of late-detected errors and repairs of early-detected errors. The confidence intervals were obtained from bootstrap analyses with 1000 replications (Davison and Hinkley, 1997; Canty and Ripley, 2013).

In the second set of analyses, the dependent variable consisted of the offset-to-repair time mentioned above (log-transformed and centered). An LMM was estimated with speakers (n=38) as single random effect, similar to the LMMs described above. The only predictor in the LMM was the detection status of the speaker's mispronunciation.

3.2 Results

3.2.1. Duration

The duration of the vowel in the repair was found to depend not only on the duration of the corresponding vowel in the reparandum (p<.0001), but also on the detection status of the speaker's error. Vowel duration in repairs of early-detected errors was on average about 15 ms longer than vowel duration in repairs of late-detected errors [p=.0131; 95% CI (+3,+36) ms]. There was no effect of offset-to-repair time on the vowel duration in the repair, nor of its interaction with detection status.

3.2.2. Maximum pitch

Preliminary analyses of the pitch differences indicated n=5 suspect cases in which the pitch measurements may have been unreliable, as suggested by a large discrepancy between maximum and average pitch (over 6 semitones discrepancy) or by a large jump in maximum pitch or in average pitch between reparandum and repair (over 9 semitones). These suspect cases were ignored for the LMMs of maximum pitch and of average pitch, leaving n=58 cases from 35 speakers.

Not surprisingly, the maximum pitch of the vowel in the repair is correlated with the maximum pitch of the same vowel in the reparandum (p<.0001); this captures effects of gender, speaker identity and of intrinsic pitch. The maximum pitch of vowels in repairs of early-detected errors and repairs of late-detected errors was found to be not significantly different [t<1, n.s.; 95% CI (−0.8,+1.0) semitones]. The (log-transformed) offset-to-repair time yielded a significant negative effect on the maximum pitch in the repair [−0.5 semitone: t=−2.31, p=.0279], so that later repairs (with shorter offset-to-repair times) tended to have a lower maximum pitch than earlier repairs (with longer offset-to-repair times). The interaction between detection status and offset-to-repair time was not significant.

[1] Due to the low number of observations, inclusion of elicitation items (n=61), as a second crossed random effect yielded unstable LMMs. Between-item variance was therefore ignored, as it was far smaller than between-subject variance for all dependent variables.

3.2.3. *Average pitch*

Just as with maximum pitch, the average pitch of the vowel in the repair is strongly correlated with the average pitch of the same vowel in the reparandum ($p<.0001$). The average pitch of vowels in repairs of early-detected errors and of late-detected errors was found to be not significantly different [t<1, n.s.; 95% CI (−1.1,+1.0) semitone]. There was no effect of offset-to-repair time on the average pitch in the repair, nor of its interaction with detection status. There was, however, a significant effect of offset-to-repair time on the average pitch of the repair (−0.7 semitone, $p=.0015$), as well as a significant interaction with detection status (+0.8 semitone, $p=.0021$). After late detected errors, average pitch in the repair tends to decrease with offset-to-repair time (i.e. later repairs of late-detected errors tended to have lower average pitch), but after early-detected errors, average pitch in the repair remains equal or increases with offset-to-repair time (i.e. later repairs of early-detected errors tended to have equal or higher average pitch).

3.2.4. *Maximum intensity*

Not surprisingly, the maximum intensity of the vowel in the repair is also strongly correlated with the maximum intensity of the same vowel in the reparandum ($p<.0001$); this captures effects of speaker identity and of intrinsic intensity. In addition, the maximum intensity of vowels in repairs of early-detected errors was found to be 3.3 dB higher than the maximum intensity of vowels in repairs of late-detected errors [$p=.0033$; 95% CI (+1.5,+5.3) dB]. There was no effect of offset-to-repair time on the maximum intensity in the repair, nor of its interaction with detection status.

3.2.5. *Average intensity*

Just as with maximum intensity, the average intensity of the vowel in the repair is strongly correlated with the average intensity of the same vowel in the reparandum ($p<.0001$). In addition, the average intensity of vowels in repairs of early-detected errors was found to be 1.7 dB higher than the average intensity of vowels in repairs of late-detected errors [$p=.0462$; 95% CI (+0.7,+4.4) dB]. There was no effect of offset-to-repair time on the average intensity in the repair, nor of its interaction with detection status.

3.2.6. *Spectral slope*

The spectral slope of the vowel in the repair is again strongly correlated with the spectral slope of the same vowel in the reparandum ($p<.0001$). The spectral slope of vowels in repairs of early-detected errors and the spectral slope of repairs of late-detected errors was found to be not significantly different [t<1, n.s.; 95% CI (−3.3,+4.4) dB]. Again there was no effect of offset-to-repair time on the average intensity in the repair, nor of the interaction with detection status, although a weak interaction effect in the predicted direction was found ($p=.1212$). After late-detected errors, spectral slope in the repair tended to decrease with offset-to-repair time (i.e. later repairs of late-detected errors tended to have a steeper or more negative spectral slope, indicating lower vocal effort), but after early-detected errors, spectral slope in the repair tended to increase with offset-to-repair time (i.e. later repairs of early-detected errors tended to have a less negative spectral slope, with more energy in the higher frequencies, indicating higher vocal effort).

3.2.7. *Offset-to-repair time*

The offset-to-repair time in repairs of early-detected errors was found to be shorter (114 ms, after back-transformation) than in repairs of late-detected errors [268 ms, after back-transformation; $p=.0322$; 95% CI (−1.18, −0.21) log ms units, corresponding to a backtransformed difference of (−186,−52) ms]. It should be noted that if we would not have focused on offset-to-repair times, but rather on error-to-repair times the difference in delay between repairs of early- and late-detected errors would have been considerably greater. The average difference in speaking time after the error and before speech is stopped is in the order of 500 ms This would bring the difference in delay in the order of 650 ms.

3.3. Summary of the acoustic measurements

In summary, we see that the phonetic properties of the vowel in the reparandum and of the same vowel sound in the repair are strongly correlated. In addition, there are significant effects of detection status (late-detected vs early-detected). Compared to repairs of late-detected errors, the vowels in repairs of early-detected errors have a longer duration, higher maximum intensity and higher average intensity. In the repairs of early-detected errors, the spectral slope tends to be less negative as the offset-to-repair time of these early-detected errors is longer. Early-detected errors yield significantly shorter offset-to-repair times than late-detected errors. We will come back to these differences in the discussion.

4. A listening experiment

In order to confirm that the acoustic differences between repairs of early- and late-detected errors can be used by the speakers as signals to their listeners, we should at least demonstrate that these differences have audible consequences. To this end we have set up a listening experiment, in which listeners were presented with pairs of excised initial CV fragments taken from both early- and late-detected speech errors. In this case we started with all 94 speech errors listed in Table 1.

4.1. Stimuli

Stimuli were prepared from the set of repaired speech errors represented in Tables 1-3. Each stimulus consisted of a pair of CV-fragments, excised with the help of PRAAT from the reparandum and from the repair. For each stimulus with the order of CV-fragments reparandum-repair there was also a counterpart stimulus with the same CV-fragments with the order repair-reparandum. This was done to neutralize potential systematic effects of the order of presentation of the 2 CV-fragments constituting one stimulus. The silent interval between offset of the first fragment and the onset of the second fragment was fixed at 250 ms. Because intensities stemmed from the original recordings and because the speakers in the speech error elicitation experiment varied widely in their vocal effort, the stimuli varied widely in loudness. The differences in loudness were such that the experiment hardly could have been run with the original overall intensities. Therefore the overall intensity of each stimulus, consisting of two successive CV fragments, was set at 70 dB above threshold, while preserving the relative intensities of the two CV fragments. In five cases the subjective sound quality was so poor that it was nearly impossible to perceive the identity of the speech sounds. These stimuli were removed from the set, reducing the number of stimuli from 2 × 94 to 2 × 89 of which 2 × 35 were detected late and 2 × 54 were detected early.

4.2. Participants

There were 11 listeners, recruited from the participant data base of UiL OTS, all students of Utrecht University. They were all native speakers of Dutch. Age ranged from 19 to 31. Listeners had no self-reported hearing deficiency.

4.3. Procedure

All listeners were tested separately in a sound-proofed booth. They had a little box with a red push button on the left and a blue push button on the right. They were told that they were participants in an experiment to study how fast people can react to a difference in loudness between brief speech fragments, and that each stimulus consisted of two such fragments. They were instructed to push the left (red) button as fast as possible when the first of the two fragments was loudest and the right (blue) button when the second of the two fragments was loudest. They were instructed to guess when they perceived the two speech fragments as equally loud. They were informed that each following stimulus would come automatically after one second and were urged to always react as fast as possible by pressing one of the two push buttons. Each individual session started with 10 arbitrarily chosen stimuli as an exercise. After that the listener could ask questions about the task. Then the series of 178 stimuli started. On a screen the listener could see a number indicating how many stimuli were still to come. The experiment took less than 10 minutes for each listener.

4.4. Results

4.4.1. Binary response

Data from 2 listeners were discarded because their miss rate was well over 50%. The miss rate of the remaining 9 listeners was 11% on average. For each presentation, listeners' response was recoded to "reparandum louder" or "repair louder", and these binary responses were analyzed by means of generalized linear mixed-effects models (GLMM; Quené and Van den Bergh, 2008), similar to a mixed-effects logistic regression. The odds of the "repair louder" response constituted the dependent variable, and speakers (n=49), listeners (n=9) and speech errors (n=89) were included as random effects. (The number of speakers is higher than in the acoustic analyses reported above, because speech errors involving different vowels in reparandum and repair were not excluded as stimuli in this perception experiment.). GLMM estimates were again obtained using the package *lme4* (Bates, Maechler and Bolker, 2013) in *R* (R Core Team, 2013). Predictors in the GLMM were the detection status of the speaker's mispronunciation (dummy coding, codes 0=late, 1=early), and an indicator of the position of the repaired speech error (dummy coding, codes 0=initial or final consonant, 1=vowel). These predictors were not included in the random part of the GLMM because then the models' terms could not be estimated properly.

The resulting GLMM shows a main effect of detection status (Z=3.689, $p<.0002$). For late-detected *consonant* errors, the log odds of a listener responding "repair louder" are −0.3438, corresponding to 41% of valid responses. For early-detected consonant errors, however, the log odds of "repair louder" responses are significantly higher at +0.8843, or 71% of the valid responses. For late-detected *vowel* errors, the log odds of "repair louder" responses are −1.6495 (or 16%) and +0.5454 (or 63%), respectively, for late-detected and early-detected vowel errors. The main effect of the position of the error (consonant vs. vowel) was also significant (Z=−2.626, p=.0044). The effect of detection status may be somewhat larger for vowel errors than for consonant errors, as shown by the weak tendency towards an interaction effect (Z=1.546, p=.1222).

4.4.2. Response times

Listeners' response times (measured from the offset of the stimulus) were analyzed by linear mixed models (LMM), using the same predictors and the same random effects as for the binary responses presented above, with MCMC estimation of significance levels (Baayen, 2011). The resulting LMM shows a marginally significant main effect of detection status (beta=−25, t=−1.711, p_{MCMC}=.0852). For late-detected errors, the estimated response time was 524 ms, whereas for early-detected errors it was 499 ms. The main effect of error position was not significant (beta<1, t<1, n.s.), nor was the interaction of the two fixed predictors (beta=+22, t<1, n.s.).

4.5. Summary of the listening experiment

In summary, we see that the detection status (late-detected vs early-detected) has a significant effect on the odds of the repair being judged as subjectively louder than the reparandum, and a marginally significant effect on the response time of that binary choice. For late-detected errors, the repair is judged louder in about 41% (16% for vowel errors), whereas for early-detected errors the repair is subjectively louder in about 71% (63% for vowel errors). Loudness judgments for early-detected errors are marginally faster, and hence subjectively easier, than for late-detected errors, for both vowel and consonant errors.

5. Discussion

It has been convincingly argued in the past that speakers who make a segmental speech error either detect this error in inner speech, before speech initiation (cf. Blackmer and Mitton, 1991; Nooteboom, 2005; Nooteboom and Quené, 2008; 2013b) or they detect the error after speech initiation, probably in overt speech or perhaps during articulation.

When a speech error is detected in inner speech, this does not necessarily imply that the error will be suppressed before speech is initiated. Apparently stopping speech that was already planned takes time, perhaps to plan a repair (Seyfeddinipur, Kita and Indefrey, 2008) and therefore, although the error was detected before speech is initiated, in many cases the speech is stopped only after speech initiation, giving rise to early interruptions of the type ***boo..good beer***. We interpret the above acoustic and perceptual results as indicating that speakers have a tendency to distract the listeners' attention from such early detected errors in three ways, viz. (a) by stopping their inadvertently initiated speech as rapidly as possible, (b) by making a repair as soon as possible and (c) by speaking the repair with more vocal effort than the reparandum. From the assumption that speakers speak the repairs of early-detected errors as fast as possible, one might have suspected that they would also speak these repairs more rapidly than the repairs of late-detected errors. We have seen, however, that the vowel durations in speech fragments taken from repairs of early-detected errors are not shorter but longer than those in speech fragments taken from late-detected errors. We suspect that this is an artifact of the time pressure induced by the task: The total time available for speaking a word pair and making a repair was limited to 2000 ms. The error-to-repair time was on average roughly 650 ms longer in late-detected errors than in early-detected errors. This considerably longer processing time before a participant initiated a repair a of a late-detected error meant that in those cases participants were more often pressed for time than in repairs of early-detected errors. This may explain why vowel durations in repairs of late-detected errors were shorter than those in repairs of early-detected errors.

The increased vocal effort in repairs of early detected errors as compared to repairs of late detected errors, although statistically significant, is far from consistent. There are many repairs of early detected errors in which the vocal effort is not greater but smaller than the vocal effort in the reparandum, and there are also many cases in which the vocal effort in the repair of late detected errors is greater, not smaller than the vocal effort in the reparandum. This inconsistency could have resulted at least in part from our definition of "early-" and "late-detected": We have classified all cases in which the reparandum was completed as "late-detected". However, in this way we may have misclassified repairs of perseverations, particularly those on the C_2, because there the error only occurs on the last segment of the word pair. If such errors were detected in inner speech by the speaker, they were nevertheless classified as "late detections" in our analysis. This may explain part of the inconsistency found. This tentative explanation was explored by a re-analysis of the perceptual results of a subset of cases, after excluding cases in which a perseveration was elicited, and in which an error on the C2 was elicited. However, a re-analysis of the perceptual results for this smaller subset (of n=29 errors, of which 12 late- and 7 early-detected) yielded effects that were similar in size and direction to those reported above for the full data set, although these effects were no longer significant due to a lack of statistical power: in the smaller subset, the early~late contrast yielded p=.0904 in response rates, and p=.1464 in response times. Thus the inconsistencies in the (acoustic and) perceptual differences between early-detected and late-detected errors cannot easily be attributed to our definition of these two categories.

Cutler (1983) and Shattuck-Hufnagel and Cutler (1999) have claimed that "prosodic marking" (here interpreted as "increased vocal effort") of the repair is more frequent for repaired lexical speech errors than for repaired segmental speech errors. Of course, in our speech error elicitation experiment we have not elicited lexical errors. The observation by Cutler and Shattuck-Hufnagel may well be correct. However, on the basis of what we find for segmental errors we would expect that the frequency of increased vocal effort would be far greater for early detected and interrupted lexical errors than for late detected lexical errors. But note that in normal spontaneous speech early interrupted lexical errors are extremely rare (Nooteboom, 2005a), although they seem to be rather frequent in the network description task used by Levelt (1983) and re-analyzed by Levelt and Cutler (1983). (Presumably this is caused by the rather frequent use in the network description task of adjectives with only a single or only a few alternatives such as *horizontal* vs *vertical* and *orange* vs a few alternative color names). The current data convincingly show that prosodic marking of the repair is not limited to lexical errors, as has been claimed by Cutler (1983) and Shattuck-Hufnagel and Cutler (1999).

With respect to those repairs that are not prosodically marked, it has been suggested that these repeat the prosody of the reparandum. This would help listeners to know what the reparandum is and to replace it with the repair (Levelt and Cutler, 1983). If this is indeed the case, one would suspect that there are quite a number of cases where the prosody of reparandum and repair is identical or at least sounds so similar that the prosody of the repair sounds like a copy of the prosody of the reparandum. In our data this would more often be the case for late-detected than for early-detected errors. Obviously, this is not so for many repairs of late-detected errors. We have seen that some 40% of late-detected consonant errors are prosodically marked in the same way as early-detected errors are prosodically marked: Vocal effort is greater in the repair than in the reparandum. Yet, on average for all late-detected errors together vocal effort is significantly less (average and maximum intensity lower and perceived loudness lower) in repairs than in reparandums. This seems to leave not too much room for the prosody of repairs being a copy of the prosody of reparandums. Indeed, informal listening to the stimuli tells us that of the 35 late-detected repaired errors used in the listening experiment, there are only 2 or 3 cases where intensity and pitch of reparandum and repair are virtually indistinguishable. In all other cases vocal effort is either higher (some 40% of the cases) or lower (the majority of cases) in the repair than in the reparandum. However, it should also be noted that in the SLIP experiment in which speech errors were elicited, there was little need for signaling what the reparandum is. The reparandum is virtually always the whole CVC CVC sequence. This is often very different in spontaneous speech or more continuous speech elicited in Levelt's (1983) network description task. There the need for prosodic signaling of what the reparandum is would be much greater than in the SLIP experiment from which our materials were harvested.

It seems that speakers have a tendency to lower their vocal effort in speaking repairs of late-detected errors. Possibly they do this to keep the error within the listeners' attention, at the same time signaling by making a repair that the error should be rejected and replaced by the repair. This ties in with an observation by Nooteboom (2010). He reported that repairs of late-detected errors are far more frequently accompanied by editing expressions such as *uhh, sorry, no, dunno (don'tknow)* etc. than repairs of early-detected errors. Such editing expressions seem to signal to the listener that an error has been made and that the error should be replaced by the repair. This does not happen with repairs of early-detected, interrupted errors. Speakers rather attempt to have these early-detected error fragments rapidly overruled by loudly spoken repairs without editing expressions.

We conclude from the current investigation that speakers follow different communicative strategies following early- and late-detected segmental speech errors. After early-detected speech errors they hasten to overrule the interrupted error with a rapid and louder repair, after late-detected speech errors they tend to signal to the listener that an error has been made and that it is being repaired.

References

Baars, B. J., Motley, M. T., and 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. (2011). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.4.

Baayen, R.H., Davidson, D.J. and 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. and Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-2*. Available: http://CRAN.R-project.org/package=lme4.

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

Boersma, P., and Weenink, D. (2009). *Praat: Doing phonetics by computer (Version 5.1.05)* [Computer program]. Available: <http://www.praat.org/>.

Canty, A. and Ripley, B. (2013). *Boot: Bootstrap R (S-Plus) Functions. R package version 1.3-9.*

Cutler, A. (1983). Speaker's conceptions of the functions of prosody. In: A. Cutler and D.R. Ladd (Eds.), *Prosody: Models and Measurements*. Heidelberg: Springer, pp. 79–91.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge. ISBN 0–521–57391–2.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*. Doctoral Thesis, Utrecht: LOT.

Goffman, E. (1981). Radio talk. In: E. Goffman (Ed.). Forms of Talk. Oxford: Blackwell, pp.197–327.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Humphreys, K. (2002). *Lexical bias in speech errors.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

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

Levelt, W.J.M. and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics, 2*, 205-217.

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, and A. Zondervan (Eds.), *The linguistic enterprise: From knowledge of language to knowledge in linguistics*, Amsterdam: John Benjamins Publishing Company, pp. 213–235.

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*, New York: Academic Press, pp. 87–95.

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

Nooteboom, S. G. (2005b). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication, 47,* (1–2), 43–58.

Nooteboom, S. G., and 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.

Nooteboom, S. G., and Quené, H. (2013a). Heft lemisphere: Exchanges predominate in segmental speech errors. *Journal of Memory and Language, 68*, 26–38.

Nooteboom, S. G., and Quené, H. (2013b). Parallels between self-monitoring for speech errors and identification of the misspoken segments. *Journal of Memory and Language, 69*, 417–428.

Plug, L. (2011), Phonetic reduction and informational redundancy in self-initiated self-repair in Dutch. *Journal of Phonetics, 39*, 299–297.

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

Quené, H., and 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 Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Available: http://www.R-project.org/

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

Shattuck-Hufnagel, S. (1987). The role of word onset consonants in speech production planning: New evidence from speech error patterns. In: E. Keller and M. Gopnik (Eds.), *Motor and Sensory Processed of Language*, Hillsdale (New Jersey) and London: Lawrence |Erlbaum Associates, Publishers, pp. 17–51.

Shattuck-Hufnagel, S. and Cutler, A. (1999). The prosody of speech error corrections revisited. In: J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A.C. Bailey (Eds.), *Proceedings of the XIVth International Congress of. Phonetic Sciences*. University of California, Berkeley, pp. 1483–1486.

Sluijter, A.M.C. and Van Heuven, V.J. (1996). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America, 100*, 2471–2485.

Van Heuven, V.J. and Sluijter, A.M.C. (1996). Notes on the phonetics of word prosody; In: Goedman, Van der Hulst, Visch (Eds., *Stress patterns of the world, part 1: Background, HIL Publications* (vol. 2), Holland Academic Graphics, The Hague, pp. 233–269.