Why are some speech errors detected by self-monitoring "early" and others "late"?

Sieb Nooteboom and Hugo Quené
Utrecht University, Utrecht, The Netherlands

Abstract

In this paper we attempt to answer the question why in self-monitoring some segmental speech errors are detected in internal, some in external speech, and others not at all. This was done by re-analyzing data obtained in two earlier published SLIP experiments. It is hypothesized that detection of errors that are similar to the correct target takes longer than detection of errors that are dissimilar. It is also hypothesized that the time available for error detection in internal speech and for detection at all is limited. Results show that indeed a major factor is the strength of phonetic contrast between two competing response candidates.

Introduction

Errors of speech can be detected by self-monitoring internal (early) or external speech (late; cf. Levelt, Roelofs & Meyer, 1999; Hartsuiker, Kolk & Martensen, 2005). It has been demonstrated that this leads to a bimodal distribution of error-to-interruption times for repaired errors, the two peaks being separated by some 450 or 500 ms (Nooteboom & Quené, 2017), confirming that detection of segmental speech errors is a two-stage process. The bimodal distribution of log interruption times can be described as two overlapping gaussians, by applying an uninformed gaussian mixture model (Fraley & Raftery, 2002; Fraley et al., 2012). This allows us to fit a separation value for the two gaussians. All interruption times below this separation value are assigned to "early detections", all longer interruption times are assigned to "late detections". An example of an early detection is the repaired error K.. PAF KIEP, an example of a late detection is the repaired error KAF PIEP.. PAF KIEP. It is assumed that early detected repaired errors are detected in internal speech, i.e. before speech initiation, and that late detected errors are detected in external speech, i.e. after speech initiation (cf. Nooteboom & Quené, 2017). In this paper we attempt to find out why some errors are detected early (i.e. internally) and others late (i.e. externally). Speech errors may also remain unrepaired, assumedly because they were not detected, for example KAF PIEP instead of PAF KIEP.

We hypothesize (1) that detection of segmental speech errors depends on comparing the sound forms of competing simultaneously active response candidates from onset to offset (KAF as an error for PAF is detected by comparing the planned sound form KAF with the simultaneously active form PAF), (2) that detection of errors similar to the correct target takes more time than detection of errors that are more dissimilar, (3) that the time available for detection in internal speech is limited, (4) that, if this time is exceeded, the error will be passed on to self-monitoring overt speech, and (5) that if the speech error also exceeds the time available for detection in overt speech, it will remain undetected. We thus distinguish between early detected, late detected and undetected speech errors.

From this account of self-monitoring internal and overt speech for segmental speech errors we derive the following predictions:

1) There are relatively more dissimilar speech errors than similar speech errors detected internally.
2) There are relatively more similar speech errors than dissimilar speech errors detected externally.
3) There are relatively more similar than dissimilar errors that remain undetected and therefore unrepaired.

An interesting question is how similarity is to be assessed. In the literature on speech errors we find examples of assessing similarity by counting distinctive features (Nooteboom, 1967; Dell, 1986). However, Guenther (2016, chapter 1) proposed that during speech production, specification of speech sounds may be different at different levels of representation. He suggested that at least the following specifications are involved in speech planning: (1) abstract phonemes (2) targets in auditory perceptual space (involved in early planning of articulation) and (3) speech motor commands (involved in specifying articulatory gestures).

So now we are confronted with two questions to be answered in this paper: (1) Is it correct that the strength of the contrast between error and correct response candidates determines whether an error is detected early or late (or not all)? (2) If so, is the contrast more phonological, to be assessed by counting distinctive features, or more phonetic, to be

https://doi.org/******/******

assessed by determining the relative strength of the contrast?

The first question will be answered by comparing frequencies and repair frequencies of errors involving a single distinctive feature (similar), viz. place or mode of articulation, with those of errors involving at least two distinctive features (dissimilar), viz. place plus mode of articulation. This will be done in Experiment 1. The second question will be answered by comparing error frequencies and repair frequencies of segmental speech errors involving (a) voicing errors in word initial stop consonants, (b) similar errors as defined for Experiment 1, (c) dissimilar errors as defined for Experiment 1, and (d) vowel errors. Voicing of consonants in word initial position is a relatively weak contrast in Dutch (Van Alphen & McQueen, 2006). Voiced and corresponding unvoiced initial stop consonant are distinguished in Dutch by the length of prevoicing. Unvoiced consonants are not aspirated. Phonetically voicing contrast is strengthened by an additional contrast in force of articulation, traditionally captured by fortis or tense for voiceless consonants and lenis or lax for voiced consonants. Vowel oppositions are supposed to provide a relatively strong contrast in Dutch.

Experiment 1

Experiment 1 has been reported as Experiment 1 in Nooteboom and Quené (2017). It was originally set up to investigate temporal aspects of detecting and repairing segmental speech errors in a SLIP (Spoonerisms of Laboratory Induced Predisposition, cf. Baars, Motley & MacKay, 1975) experiment. Here we limit description of Experiment 1 to those aspects that are relevant to the current task.

Method of Experiment 1

There were 106 participants. Interactive segmental speech errors were elicited by having CVC CVC Dutch word pairs (stimulus items), each preceded by 5 CVC CVC word pairs, the last three of which triggered a reversal of the two word initial consonants, as in BOUW JOOL, LIJF DEED, KEN PIT, KOET POP, KAS PIET, preceding the stimulus word pair PAF KIEP. There were two stimulus lists. In each list there were 32 stimuli, 16 with the two initial consonants differing in a single distinctive feature (place or mode of articulation; similar), and 16 with the two initial consonants differing in two distinctive features (place plus mode of articulation; dissimilar).

There were 23 filler stimuli, preceded by 0, 1, 2 3 or 4 CVC CVC word pairs not triggering a segmental reversal. After each test stimulus and each filler stimulus a sequence of "?????" was presented, as a cue to speak aloud the last word pair seen. After the "?????" there followed a presentation of the Dutch word for "repair?", to elicit sufficient repairs.

Each speaker was tested individually in a sound-treated booth. Presentation of precursors, stimuli, and "repair" cues always lasted 900 ms followed by a blank interval of 100 ms. Responses were categorized using Praat (Boersma & Weenink, 2016) for the current purpose as (0) fluent and correct, (1) hesitations and omissions, (2) completed and interrupted single elicited segmental errors, (3) completed and interrupted single other errors, (4) completed and interrupted multiple other errors. The current analysis focuses mainly on category 2.

Results of Experiment 1

A first breakdown of the observed speech errors is given in Table 1.

Table 1. Numbers of responses, broken down by response category and repair status.

response category	repair status		total
subjects	not repaired	repaired	
fluent & correct	5821	0	5821
hesitations & omissions	64	40	104
single elicited errors	298	115	413
single other errors	187	31	218
multiple errors	192	36	228
total	6562	222	6784

In our further analyses we will mainly focus on the 413 single elicited, errors, i.e. those errors the SLIP technique was meant to elicit. Of these 298 were not repaired, presumably because they were not detected by self-monitoring either in internal speech or after speech initiation, and 115 were repaired before or after speech initiation.

Speech errors were either "detected early", or "detected late", or else remained unrepaired, supposedly "undetected". The categorization as to whether a specific repaired error was detected "early" or "late" was made on the basis of an uninformed gaussian mixture model applied to "error-to-interruption times" (cf. Fraley & Raftery, 2002; Fraley et al., 2012). The resulting bimodal distribution is shown in Figure 1.

A relevant independent variable is "similar" vs "dissimilar" interacting consonants in the error; a relevant dependent variable is whether the error was detected "early" or "late" or "not detected" at all. This breakdown is given in Table 2.

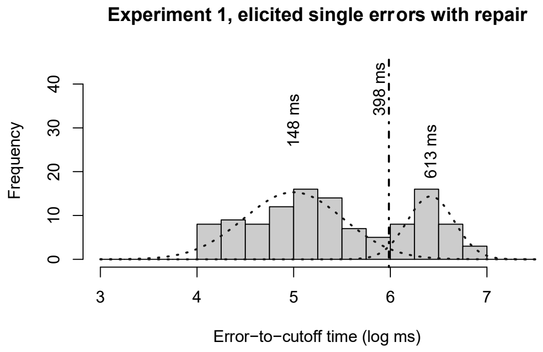

Figure 1. Histogram of log-transformed durations of error-to-interruption intervals, for N = *114 repaired errors (the error-to-interruption interval of 1 repaired error was missing). Dotted lines indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicated the interpolated boundary value.*

Table 2. Numbers of single elicited segmental errors, broken down by "similar" vs "dissimilar" and by "early detected" vs "late detected" vs "not detected".

	Detection			
error category	early	late	not	Total
similar	26	18	192	236
dissimilar	54	17	166	177
total	80	35	298	413

Here we see that "similar" errors are far more frequent than "dissimilar" errors. This difference was found to be significant in a Bayesian binomial logistic mixed-effects regression (GLMM; Bürkner, 2017, 2018), with errors as hits and participants as random intercepts, and similarity as fixed effect; the response category "not detected" was used as baseline. The data reported in this paper, and full details of all statistical analyses, are available at https://osf.io/gxjnm/. Log odds are on average −2.79 for items eliciting interaction among "similar" consonants (baseline) and they are lower by −0.32 for items involving "dissimilar" consonants (with 95% highest posterior density interval [−0.52, −0.12]), suggesting a significant difference between similar and dissimilar.

The three-way classification of errors as "early detected", "late detected" and "not detected" (Table 2) was further analyzed with another Bayesian multinomial mixed effects regression model, with participants as random effect, contrast as fixed predictor (with similar interacting consonants as baseline). The odds of an error being detected early are indeed far lower for similar items (posterior mean −2.06) than for dissimilar items (mean −0.71, with non-overlapping posterior density intervals. The difference in (very low) odds of late detections between "similar" and "dissimilar" was not found to be significant.

Discussion of Experiment 1

The results of Experiment 1 demonstrate that strength of contrast determines whether segmental speech errors are detected in internal speech or not. Of course, in the current analysis, strength of contrast was expressed by counting distinctive features. In Experiment 2 we attempt to find out whether strength of contrast should rather be expressed phonetically.

Experiment 2

Experiment 2 has been reported as Experiment 2 in Nooteboom and Quené (2017). Experiment 2 is largely identical to Experiment 1, but in addition to stimuli eliciting interactions between "similar" and "dissimilar" consonants, we also added a category of stimuli eliciting interactions between "voiced" and "unvoiced" consonants and a category of stimuli eliciting interactions between the vowels of the two CVC words. Here we limit description of Experiment 2 to aspects that are relevant to the current task.

Method of Experiment 2

There were 124 participants. There were two stimulus lists. In each list there were 32 stimuli eliciting interactions between word initial consonants differing in place and/or mode of articulation, of which 16 differing in a single distinctive feature (place or mode of articulation; similar), and 16 with the two initial consonants differing in two distinctive features (place plus mode of articulation; dissimilar). There were also 16 stimuli eliciting interactions between voiced and unvoiced word initial consonants and 16 stimuli eliciting interactions between the vowels of the two CVC words. There were also 46 filler stimuli, with a number of precursors varying between 0 and 4. The precursors of the fillers did not prime interactions. Further details of the materials, the procedure and the scoring were the same as in Experiment 1. The current analysis focuses mainly on category 2.

Results of Experiment 2

A first breakdown of the numbers of single elicited errors is given in Table 3.

Speech errors were either "detected early", or "detected late", or else remained unrepaired, supposedly "undetected". The categorization as to whether a specific repaired error was detected "early" or "late" was made on the basis of an uninformed gaussian mixture model applied to "error-to-interruption times" (cf. Fraley & Raftery, 2002; Fraley et al., 2012). The resulting bimodal distribution is shown in Figure 2.

As is clear from Table 4, the numbers of total errors are conspicuously different for the four classes of stimuli. By far the most errors are made against "voicing", which confirms that the voicing contrast is relatively weak in Dutch. By far the fewest errors are made against "vowels", which confirms that the contrast between vowels is relatively strong in Dutch. These differences were again analyzed in a Bayesian GLMM (Bürkner, 2017, 2018), with errors as hits and participants as random intercepts, and similarity as fixed effect (see https://osf.io/gxjnm/ for details). Log odds are on average −2.72 for items eliciting interaction among "similar" consonants (baseline) and they are lower by −0.71 logits (95% HPDI [−1.02, −0.43]) for items involving "dissimilar" consonants, again suggesting a significant difference between similar and dissimilar consonants. Moreover, the log odds of errors involving consonantal voicing contrast are higher than the baseline by +0.91 logits (95% HPDI [0.72, 1.10]), and the log odds of errors involving vowels are lower by −0.96 logits (95% HPDI [−1.28, −0.66]).

The three-way classification of errors as "early detected", "late detected" and "not detected" (Table 4) was further analyzed with another Bayesian multinomial mixed effects regression model, with participants as random effect, contrast as fixed predictor (with similar interacting consonants as baseline). The only significant effect was that the log odds for early detection were significantly lower for voicing errors (30:532) than for similar errors (45:214; the difference being −1.36 logits, 95% HPDI [−1.95, −0.78]).

Table 3. Numbers of responses, broken down by response category and repair status.

response category	repair status		total
	not repaired	repaired	
fluent & correct	13069	0	13069
hesitations & omissions	228	67	295
single elicited errors	956	184	1140
single other errors	570	35	605
multiple errors	473	34	507
total	15296	320	15616

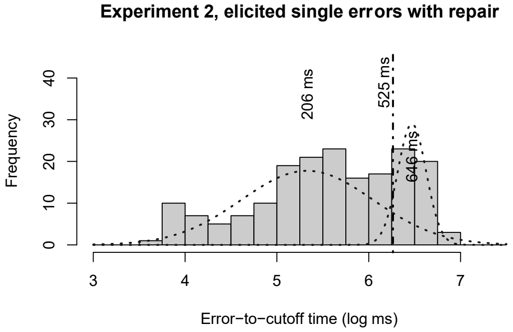

Figure 2. Histogram of log-transformed durations of error-to-interruption intervals, for N = 182 repaired errors (error-to-interruption intervals of 2 repaired errors were missing). Dotted lined indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicated the interpolated boundary value.

Table 4. Numbers of single elicited segmental errors, broken down by error category (see text) and by detection status: detected "early" or "late" or "not detected".

	detection			
error category	early	late	not	total
voicing	30	16	532	579
similar	45	16	214	275
dissimilar	38	7	125	171
vowels	23	7	85	115

Discussion of Experiment 2

Experiment 2 did not replicate the significant difference in error detection between errors involving similar and dissimilar consonants. Apparently, results were somewhat noisier than in Experiment 1. However, the significant difference between voicing errors and similar errors, both error categories involving a contrast of a single distinctive feature, suggests that speech errors are detected on the basis of more phonetic than phonological contrast (see general discussion below). We had also predicted a significant difference in frequency of early detection between dissimilar errors and vowel errors. That this effect did not show up possibly is due to the circumstance that detection of vowel errors in Dutch takes considerably more time than detection of consonant errors. This is so because the first part of Dutch long vowels and Dutch diphthongs sound as a corresponding Dutch short vowel.

General discussion

The main question we have attempted to answer in the present investigation is: "Why are some segmental speech errors detected by self-monitoring in internal speech, others in external speech, and others not at all?" The results of Experiment 1 demonstrate that a major factor is the strength of contrast between two competing response candidates, as assessed by the relative frequency of error commitment: Detection of segmental speech errors involving a weak contrast takes more time than detection of segmental speech errors involving a stronger contrast. The time available for detection in internal speech, before speech initiation, is limited. If this time is exceeded for a particular speech error, detection is likely to be postponed to a later stage of speech preparation, where articulation is initiated. In case also the time needed for detection of an error at this later stage of the speaking process is exceeded, the error remains undetected and unrepaired.

We have also attempted to find out whether contrast between competing segments involved in error detection by self-monitoring is phonological, i.e. in terms of number of distinctive features, or rather phonetic. Results of Experiment 2 suggest that contrast on the levels of representation where segmental errors are detected by self-monitoring is phonetic, to be specified in terms of more gradient segmental properties such as auditory perceptual contrast or articulatory contrast. The evidence for this conclusion in Experiment 2 stems from the comparison in frequency of "early" detection between errors involving the weak contrast in voicing and errors involving the stronger contrast in place or mode of articulation.

We wish to point out that if we apply the feature system proposed by Chomsky and Halle (1968), there sometimes is a major difference in terms of distinctive features between our errors against similar consonants in Experiment 1 and voicing errors in Experiment 2. For example, place of articulation of /t/ is specified by two distinctive features, viz. +coronal and +anterior (to distinguish labiodental /t/ from palatal /c/ that is +coronal and −anterior), whereas voicing is always specified by a single distinctive feature. This feature system brings phonology somewhat closer to phonetics. However, traditionally voiceless and voiced consonants were also assigned the features fortis and lenis or tense and lax, over and above the presence or absence of voicing. The strength of contrast between similar consonants differing in place or mode of articulation on the one hand and voicing on the other cannot be captured by counting features in some reified abstract phonological feature system, as done for example in Ulicheva et al. (2021).

The current results fit into an account of speech planning and self-monitoring for speech errors with different stages of planning. We follow Guenther (2016) in assuming that articulatory movements are planned internally in terms of sequences of targets in auditory perceptual space. During this stage segmental errors are detected "early". About 450 or 500 ms later, these segments are transformed into articulatory gestures, specified in terms of temporally coordinated motor commands. During this stage "late" error detection occurs.

References

Baars, B. J., M. T. Motley & D. G, MacKay. 1975. Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 14(4): 382–391. https://doi.org/10.1016/S0022-5371(75)80017-X

Boersma, P. & D. Weenink. 2016. Praat: Doing phonetics by computer (version 6.0.19). https://www.praat.org/ (accessed 24 January 2021).

Bürkner, P.C. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.C. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1): 395–411.

Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Dell, G. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3): 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Fraley, C. & A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631. https://doi.org/10.1198/016214502760047131

Fraley, C., A. E. Raftery, T. B. Murphy & L. Scrucca. 2012. mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Report No. 597, Department of Statistics, University of Washington.

Guenther, F. H. 2016. *Neural Control of Speech*. Cambridge, MA: The MIT Press.

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

Levelt, W. J. M., A. Roelofs & A. S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Nooteboom, S. G. 1967. The tongue slips into patterns. In: A. Sciarone, A. J. van Essen, A. A. van Raad (eds.), *Nomen, Leyden Studies in Linguistics and Phonetics*, The Hague: Mouton, 114–132.

Nooteboom, S. G. & H. Quené. 2017. Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language* 95, 19–35. https://doi.org/10.1016/j.jml.2017.01.007

Ulicheva, A., K. D. Roon, Z. Cherkasova & P. Mousikou. 2021. Effects of phonological features on reading-aloud latencies: A cross-linguistic comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. http://dx.doi.org/10.1037/xlm0000893.

Van Alphen, P. M. & J. M. McQueen. 2006. The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human*

Perception and Performance 32(1). 178–196. https://psycnet.apa.org/doi/10.1037/0096-1523.32.1.178