SELF-MONITORING OF SPEECH ERRORS: EFFECTS OF PHONETIC CONTRAST

H. Quené and S.G. Nooteboom

Institute for Language Sciences, Utrecht University, the Netherlands
h.quene@uu.nl, s.g.nooteboom@uu.nl

ABSTRACT

Speech sound errors are known to occur more often as the interfering speech sounds are phonetically more similar. This paper aims to test the general hypothesis that the phonetic contrast between interfering speech sounds also mediates the odds of detection and repair of such a speech error. This was investigated by re-analyzing responses from four published experiments in which speech errors had been elicited, here using Bayesian modeling with random effects of participants and stimuli.

Results show that with increasing phonetic contrast among the segment involved, speech errors occur less frequently (as expected); those errors tend to be detected and repaired more frequently, with early repairs being more prevalent than late repairs, as predicted. These patterns suggest that repair is not triggered by conflict during production, but by error detection during self-monitoring.

Keywords: self-monitoring, speech error, speech production, speech preparation.

1. INTRODUCTION

Most speakers produce occasional speech errors, some of which they detect and repair. Detection (and repair) may occur relatively *early* after the error (before speech is initiated), e.g. *v..horizontal* [1]; presumably such detection is based on the speaker "hearing" her/his own internal speech, as the error-to-interruption interval is too short for auditory processing and speech interruption. Detection may also occur relatively *late* after the error, presumably based on the speaker hearing his/her own overt external speech [2]. Both routes of detection involve comprehension or perception of the speaker's own (internal or external) speech. Production-based models of self-monitoring, by contrast, presume that self-monitoring occurs during speech production (not perception). In one such model [3], conflict may arise between multiple speech sounds competing for the same slot in the sequence being prepared, and this conflict in itself (indirectly) triggers prevention or repair of the impending speech sound error.

The present paper aims to test the general hypothesis that the phonetic contrast (distance, inverse of similarity) between the competing speech segments affects self-monitoring. Speech errors are known to occur less frequently as the phonetic contrast between competing sounds increases [e.g. 4, 5], but here we focus on whether error detection and repair too is mediated by phonetic contrast among the segments involved. With increasing phonetic contrast between competing sounds, we predict (1) that errors will *occur* less frequently (as reported before), (2) that those errors will however be *detected* more frequently (due to their greater saliency during monitoring), and (3) that those detections will occur *early* (rather than late) more frequently.

2. METHOD

To test the above predictions, this paper re-analyzes data from four so-called SLIP experiments [6] involving Dutch elicited speech errors. The raw data were taken from [7, 8].

2.1. SLIP experiments

SLIP experiments aim to elicit exchange errors between the initial consonants in a two-word C_1VC C_2VC stimulus, e.g. *zaal boom* in Dutch. The exchange is elicited by means of presenting priming C_2VC C_1VC precursors, e.g. *bon zet, bek zeef, baai zoop*. The precursor word pairs and the stimulus word pair appear subsequently on the computer screen (for 900 ms + 100 ms blank screen) and the participant needs to read each word pair silently. Every now and then, a visual prompt (*??????*) appears on the screen for 900 ms, upon which the participant has to speak aloud the last-presented word pair. In the SLIP experiments re-analysed here, the target stimulus was always the 5th word pair in a trial, and it was always preceded by at least 3 priming precursors. A typical spoken response containing the elicited error might be *baa.. zaal boom*. Participants were under time pressure, because they had only 900 ms to respond (after the prompt) before a buzzer would sound.

Dependent variables are the category of response (see below), and for error responses whether or not the participant overtly detected his/her speech error (true for the example above), and for detected errors also the time delay (in ms) from the onset of the erroneous consonant to the moment the speaker

finished or broke off her/his spoken response (e.g. duration of *baa*). From the response set, hesitations and omissions were removed, with N=32 368 observations from 433 participants remaining.

2.2. Phonetic contrast

The SLIP stimuli in the present corpus varied in the strength of phonetic contrast (or similarity) between the two interfering initial consonants in the stimulus word pair. The two consonants could differ

- in two features, viz. both place and manner of articulation (labeled *pm2*; 15 096 observations from all experiments);
- in a single feature, viz. either in place or in manner of articulation, but not in both (labeled *pm1*; 13 469 observations from all experiments);
- in their voicing feature (labeled *voicing*; 3 803 observations from two experiments).

Although the latter two contrasts both involve a single feature, the voicing contrast has been shown to be weak for Dutch speakers [9], and Dutch listeners perceive voicing of initial consonants less accurately than either place or manner [10, p.3674].
Observations involving vowel exchanges (in one experiment) have been ignored in the present paper, because the timing of (errors in) initial consonants and nuclear vowels cannot be compared sensibly.

2.3. Responses

Responses were classified into one of the categories listed in Table 1.

label	description	example
fluent	fluent, correct	*zaal boom*
elic. error (repaired)	elicited single error, repaired	*baa.. zaal boom*
elic. error (unrepaired)	elicited single error, unrep'd	*baal zoom*
other error	other error(s)	*boog baan*
hesitation	hesitation or omission	*...eh...*

Table 1: Response categories for an example stimulus *zaal boom* in a Dutch SLIP experiment.

In total, there were 28 508 fluent responses, 1 803 elicited errors (5.6%), and 2057 other errors. The SLIP technique is obviously very inefficient in eliciting speech errors. Moreover, errors were distributed very unequally over participants, as shown in Fig.1: 63 participants (15%) did not produce any elicited error over n=72 or n=110 trials per experiment, and most participants produced only a single elicited error. The top 77 (18%) contributing participants made half of the elicited errors (dark bars in Fig.1), and the top 177 participants (or 41%) contributed 1442 (or 80%) of the elicited errors. The distribution of errors over stimuli is similarly unequal. In our analyses below we will therefore consider both the effects of participants and of stimulus items.

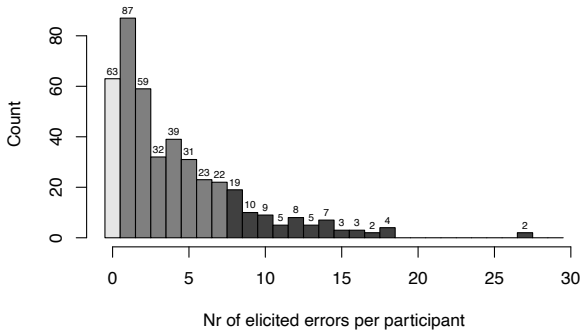

Figure 1: Distribution of the number of elicited errors, over 433 participants in the corpus of SLIP responses.

2.4. Elicited errors detected early or late

Of the 1803 elicited errors, 483 (27%) had been detected and repaired. Following [11] the bimodal distribution of the log-transformed error-to-cutoff times was used to classify repairs as either *early* (presumably based on internal speech) or *late* (presumably based on overt speech, cf. §1). Unsupervised mixture modeling [12, 13] in R [14] suggested two underlying gaussian components making up this bimodal distribution. (Here 5 responses were ignored because of missing error-to-cutoff times, and 5 were ignored because of outlier values between 7 and 15 ms.) Peaks of the two gaussians were at 133 ms and 592 ms respectively. Using the threshold error-to-cutoff time at 425 ms, errors were classified as detected and repaired *early* (371 errors, 21%) or *late* (112 errors, 6%).

2.5. Models

The three predictions above were tested by means of three mixed-effects models, viz. one for each prediction. First, using all remaining responses, multinomial model (1) assesses the effect of phonetic contrast (similarity, 3 categories, see §2.2) on the odds of an elicited error and on the odds of other error(s); the baseline response category are the fluent and correct responses.

(1) responsecat ~ 0 + sim +
(1 + sim | ID) + (0 + sim || stim),
family=multinomial, data=allresponses

The second, binomial model (2) zooms in on the 1803 elicited errors (see §2.3), and assesses the effect of phonetic contrast (similarity) on the odds of an

elicited error being detected; the baseline repair category are the undetected (unrepaired) responses.

(2) (repaircat!=unrep) ~ 0 + sim + (1 + sim | ID) + (0 + sim || stim), family=binomial, data=elicitederrors

Third, multinomial model (3) also zooms in on the 1803 elicited errors, and assesses the effect of phonetic contrast (similarity) on the odds of an early repair and on the odds of a late repair; the baseline repair category are again the unrepaired responses.

(3) repaircat ~ 0 + sim + (1 + sim | ID) + (0 + sim || stim), family=multinomial, data=elicitederrors

Thus, models (1) to (3) included population-level effects of phonetic contrast (or similarity), group-level effects of participants (ID) and of stimuli, and allowed effects of phonetic contrast to vary over participants and over stimuli ("random slopes").

2.6. Analysis

Models (1) to (3) were estimated using Bayesian methods in R [15, 16, 17]. Each model was estimated in 4 independent chains of 3000 iterations (with 1000 warmup), using NUTS sampling. This yielded 8000 post-warmup iterations.

For group-level ("random") estimates, we report the 95% credibility interval of the posterior distribution. For population-level ("fixed") estimates, we report the 95% highest posterior density interval (HDI) [18], which is the narrowest interval containing 95% of the probability mass of the posterior distribution. If two model parameters have non-overlapping HDIs, then we have good grounds to believe that those parameters are different. For all models reported below, R-hat and other diagnostics did not indicate any convergence problems.

3. RESULTS

In all models (1−3), between-item variation (in items' odds per category) was similar across conditions of phonetic contrast. In the multinomial models (1) and (3), between-participants correlations among their odds across conditions were not credibly different from zero.

Multinomial **model (1)** estimated the odds of an elicited error, and of unelicited error(s), against the baseline of correct and fluent responses (*N*=32 368 observations). The **group-level** coefficients showed that between-participant variation (in participants' odds of the elicited error and in their odds of other errors) was higher in conditions eliciting errors in place and/or contrast (i.e. in conditions *pm1* and *pm2*) as compared to the condition eliciting errors in *voicing* [for elicited errors: sd(*intercept*) 0.54 (0.42, 0.66), *pm1* +0.31 (0.06, 0.54), *pm2* +0.51 (0.26, 0.73)]. Thus, participants are more similar in their propensity of voicing errors than in their propensity of errors involving place and/or manner -- even though voicing errors were elicited in only two of the four SLIP experiments (see §2.2).

The **population-level** coefficients of model (1), for elicited errors, are illustrated in the left panel of Figure 2, broken down by phonetic contrast.

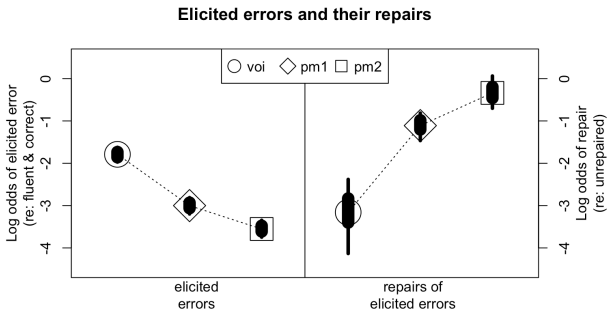

Figure 2: Summary of posterior distributions of population-level coefficients in models (1) and (2), broken down by strength of phonetic contrast. Symbols are plotted at the median of the posterior, thick lines denote 50% HDI and thin lines denote 95% HDI (see text). Left: model (1), log odds of elicited error; right: model (2), log odds of repair of elicited error.

The odds of an elicited error are highest if a voicing error is elicited, considerably lower if a place-or-manner error is elicited, and lowest if a place-and-manner error is elicited. The odds of other (non-elicited) errors, not shown in Figure 2, follow the same pattern, with medians at −2.60, −2.74 and −3.02 respectively, but with overlapping 95% HDIs.

Binomial **model (2)** zooms in on the odds of repair of the elicited errors only (*N*=1803 observations). The **group-level** coefficients showed again that between-participant variation (in their odds of repair of an elicited error) was higher in conditions eliciting *pm1* and *pm2* errors as compared to the condition eliciting voicing errors [repair: sd(*intercept*) 0.93 (0.58,1.28), *pm1* +0.60 (0.04, 1.36), *pm2* +0.70 (0.02, 1.70)]. Thus, participants are more similar in their propensity to repair an elicited voicing error than to repair an elicited error involving place and/or manner.

The **population-level** coefficients of model (2) are illustrated in the righthand panel of Figure 2. Remarkably, the pattern is the opposite of that in the lefthand panel: the odds of repair of an elicited error are lowest if a voicing error is elicited [−3.21 (−4.24, −2.44])], considerably higher if a place-or-manner is elicited [−1.12 (−1.46, −0.81)], and highest if a place-and-manner error is elicited [−0.34 (−0.72, +0.04)].

Multinomial **model (3)** also zooms in on the elicited errors only. The **group-level** coefficients showed once again that between-participant variation

(in their odds of early and of late repairs) was higher in conditions eliciting errors in place and/or contrast (i.e. in conditions *pm1* and *pm2*) as compared to the condition eliciting errors in *voicing* [early repairs: sd(*intercept*) 0.60 (0.05, 1.12), *pm1* +0.49 (0.02, 1.21), *pm2* +0.57 (0.03, 1.33); late repairs: sd 0.90 (0.13, 1.64), *pm1* +0.65 (0.03, 1.65), *pm2* +1.38 (0.09, 3.17)]. Thus, participants are more similar in their repair propensity of voicing errors than in their repair propensity of errors involving place and/or manner. Because there were only 30 *pm2* errors being repaired late, the variation among participants in estimating the odds of late repair must be very large.

The **population-level** coefficients of model (3) are illustrated in Figure 3.

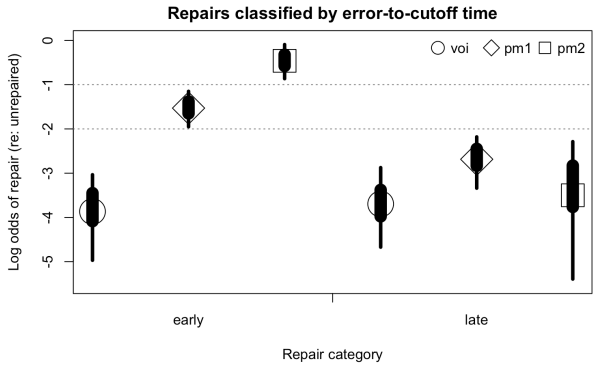

Figure 3: Summary of posterior distributions of population-level coefficients in model (3), broken down by strength of phonetic contrast and by type of repair. (See Figure 2.)

The odds of *early* repair are lowest if a voicing error is elicited [−3.93 (−4.97, −3.04)], considerably higher if a place-or-manner error is elicited [−1.54 (−1.95, −1.15)], and highest if a place-and-manner error is elicited [−0.46 (−0.86, −0.09)], with non-overlapping 95% HDIs. The odds of late repair are approximately the same across the phonetic contrasts. Thus the effects illustrated in the righthand panel of Figure 2 (odds of repair) are mainly due to early repairs, and not to late repairs.

4. DISCUSSION

The results of model (1) (Fig.2, left) confirm that the relative number of speech errors increases with phonetic similarity of the interfering consonants in the stimulus (and decreases with their phonetic contrast) [e.g. 4, 5], most clearly for interfering consonants contrasting in place and/or in manner. Moreover, the higher error rates for stimuli involving the voicing contrast confirm that this voicing contrast is indeed considerably weaker in Dutch (in speech preparation and articulation and monitoring) than either a place or a manner contrast [cf. 9].

In this paper, however, our focus is on whether and how quickly these elicited speech errors are detected and repaired. We have predicted that errors involving a stronger phonetic contrast are detected more frequently. This prediction was confirmed by the results of model (2) summarized in Fig. 2 (right). Elicited voicing errors (circles) are relatively common, but they are detected and repaired relatively rarely; elicited place-and-manner errors (squares) are relatively rare, but these are detected and repaired relatively frequently (with high odds of detection and repair); elicited place-or-manner errors have intermediate odds of error and of repair.

The results of model (3) in Fig. 3 show that the effect of phonetic contrast works mostly in early repairs, rather than in late repairs. Elicited voicing errors (circles) are hardly repaired early; elicited place-and-manner errors (squares) are repaired early quite often; elicited place-or-manner errors (diamonds) have intermediate odds of early repair, all with non-overlapping posterior distributions.

Regarded differently, the prevalence of early repair over late repair increases with phonetic contrast: there is no such prevalence for voicing errors (circles); the prevalence is strongest for repair of errors involving both place and manner contrasts (squares), with an intermediate prevalence for errors involving either a place or a manner contrast (diamonds). Thus speech errors involving a strong phonetic contrast tend to be detected and repaired early (Fig. 3); errors involving a weak contrast such as voicing tend to be detected early less often (than strong-contrast errors, Fig. 3) or these weak-contrast errors tend to be detected not at all (low odds of detection, Fig. 2).

The pattern in Figure 2 is difficult to reconcile with a conflict-based theory of self-monitoring [e.g. 3]. We assume that stronger phonetic contrast between two competing items would result in a greater separation between the two items in their relative activation and therefore in *less* conflict during production. (This would lead to fewer errors, which was indeed observed in the lefthand panel of Fig. 2.) Less conflict would however also lead to fewer (and later) detections and repairs, contrary to the observed pattern in Figs. 2 and 3.

In conclusion, phonetic contrast is a major determinant of early detection vs. late detection vs. non-detection of segmental speech errors in self-monitoring. We also conclude that "cognitive control" as a mechanism that prevents or repairs segmental errors is not triggered by conflict between competing response candidates, but rather by error detection.

5. REFERENCES

[1] Levelt, W.J.M. 1989. *Speaking: From intention to articulation*. MIT Press.

[2] Hartsuiker, R. J., Kolk, H. H. J. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology* 42(2), 113–157.

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

[4] Nooteboom, S.G. 1973. The tongue slips into patterns. In: V.A. Fromkin (ed) *Speech Errors as Linguistic Evidence*. Mouton, 144–156.

[5] Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3), 283–321.

[6] Baars, B. J., Motley, M.T. 1974. Spoonerisms: Experimental elicitation of human speech errors. *Catalog of Selected Documents in Psychology* 3, 28–47.

[7] Nooteboom, S.G., Quené, H. 2008. Self-monitoring and feedback: a new attempt to find the main cause of lexical bias in phonological speech errors. *J. Memory and Language* 58(3), 837–861.

[8] Nooteboom, S.G., Quené, H. 2013. Parallels between self-monitoring for speech errors and identification of the misspoken segments. *J. Memory and Language* 69(3), 417–428.

[9] van Alphen, P.M., Smits, R. 2004. Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. *J. Phonetics* 32(4), 455–491.

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

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

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

[13] Fraley, C., Raftery, A.E. 2002. Model-based clustering, discriminant analysis and density estimation. *J. Am. Statistical Assoc.* 97, 458, 611–631.

[14] R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

[15] Bürkner, P.-C. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Statistical Software* 80(1), 1–28

[16] Bürkner, P.-C. 2018. Advanced Bayesian multilevel modeling with the R package brms. *R Journal* 10(1), 395–411.

[17] Makowski, D., Ben-Shachar, M.S., Lüdecke, D. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *J. Open Source Software* 4(40), 1541.

[18] McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press.