Self-monitoring as reflected in identification of misspoken segments

Sieb Nooteboom & Hugo Quené

Utrecht Institute of Linguistics IuL OTS, The Netherlands

Abstract

Most segmental speech errors probably are articulatory blends of competing segments. Perceptual consequences were studied in listeners' reactions to misspoken segments. 291 speech fragments containing misspoken initial consonants plus 291 correct control fragments, all stemming from earlier SLIP experiments, were presented for identification to listeners. Results show that misidentifications (i.e. deviations from an earlier auditory transcription) are rare (3%), but reaction times to correctly identified fragments systematically reflect differences between correct controls, undetected, early detected and late detected speech errors, leading to the following speculative conclusions: (1) segmental errors begin their life in inner speech as full substitutions, and competition with correct target segments often is slightly delayed; (2) in early interruptions speech is initiated before competing target segments are activated, but then rapidly interrupted after error detection; (3) late detected errors reflect conflict-based monitoring of articulation or monitoring overt speech.

1. Introduction

Research reported in [2] and [4] on the articulation of speech errors elicited with metronome-controlled tongue twisters demonstrates that many segmental speech errors are not full, categorical, substitutions of one phoneme segment by another, but rather articulatory blends of two competing segments. In [4] the authors assume that in inner speech a segmental speech error arises when two competing segments are activated for the same slot, and that the activation of both competing segments is passed on to articulation, leading to conflicting articulatory gestures. From these findings we conclude that very likely misspoken segments carry acoustic and perceptual consequences of their discordant articulatory origin. In [9] it is argued that these properties of speech errors cannot be studied very well by means of auditory transcription, because, due to limitations of perception, articulatorily ambiguous speech sounds are either perceived as the misspoken segments or as the correct targets. However, very likely another measure of perceptual differences, such as reaction time in an identification task, would be sensitive enough to show subtle differences in perceptual clarity originating from competing articulatory gestures.

The basic idea of the research reported here is that potentially properties of self-monitoring for speech errors can be investigated indirectly through having listeners identify misspoken segments and their correct controls as elicited in SLIP experiments. Perceptual unclarity of these segments resulting from conflicting articulatory gestures would be reflected in (a) number of misidentifications and (b) average reaction times. The relation between perceptual clarity and self-monitoring can be established because of each misspoken segment it is known whether the speech error was or was not detected and repaired by the speaker, and, if detected, whether detection was early, as in *boo .. good beer* or late as in *bood geer ... uh good beer*. Because early detected errors in initial consonants very often are repaired when only part of the following vowel has been spoken, all speech sounds to be compared are to be presented in brief CV speech fragments. We have the following predictions:

(1) **Speech fragments excised from elicited segmental errors will on average have more misidentifications and longer reaction times than speech fragments from correct controls**.

This is because speech errors suffer from articulatory blending and correct controls do not.

(2) **Speech fragments excised from undetected errors will on average have more misidentifications and longer reaction times than speech fragments from detected errors.**

This prediction is based on the assumption that perceptually clear misspoken segments are easier to detect for the monitor as errors than perceptually unclear misspoken segments. This assumption follows from the comprehension-based monitor proposed in [3] plus the idea that error detection is a function of comparing the error form with the correct target form [7]. Recently a theory of production-based conflict monitoring is proposed, where the probability of error detection increases with the amount of conflict between competing segments [8]. From this theory one would make the inverse prediction.

(3) **Speech fragments excised from late detected errors will on average have more misidentifications and longer reaction times than speech fragments from early detected errors**.

The reason for this prediction is that in early detected speech errors (*boo ... good beer*) speech is initiated before the monitor has resolved the conflict between the two competing segments (cf. [7]: error detection follows speech initiation). This suggests that activation of the correct target segment often comes slightly later than activation of the error segment, and may come too late to have much effect on articulation before speech is initiated. In late detected errors both competing segments would be fully activated before activation is passed on to articulation and speech is initiated. Therefore speech will be often affected by conflicting articulatory gestures, and resulting speech sounds will suffer from perceptual unclarity,

From these three predictions it follows that we expect the following order of our four conditions in increasing number of misidentifications and increasing reaction times:

1) Speech fragments from correct controls

2) Speech fragments from early detected errors

3) Speech fragments from late detected errors

4) Speech fragments from undetected errors

2. Method

We selected 291 speech errors on initial consonants from the speech errors elicited with the SLIP technique [1] in two experiments described in [7]. The only selection criterion was that each of these speech errors had a correct control where no speech error was elicited, spoken by the same speaker. Thus we also had 291 correct controls. The 291 speech errors were either not detected and repaired by the speaker (158 cases), or early detected and repaired (as in *boo ... good beer*; 80 cases), or late detected and repaired (as in *bood geer ...uh... good beer*; 53 cases). From each speech error and each correct control a speech fragment was excised containing the initial consonant and 40 ms of the following vowel. All 582 speech fragments were presented over headphones to 21 listeners in a simple open response identification task, in which listeners were asked to press the key on a keyboard corresponding to the consonant sound they heard. Misidentifications defined as deviations (other than those stemming from misperception of voiced/voiceless) from an auditory transcription of the same material described in [7] were assessed for all 12222 responses and reaction times (RTs) were assessed for correctly identified speech fragments, and only for those stimuli that had less than 4 misidentifications. Misperceptions of voice were not counted as misidentifications because the voiced/voiceless feature in initial position is not very robust in Dutch, and misperceptions of voice in meaningless speech fragments are unpredictable.

3. Results

Figure 1 gives the relative frequencies of expected and observed misidentifications in the four conditions formed by the origin of the speech fragments.

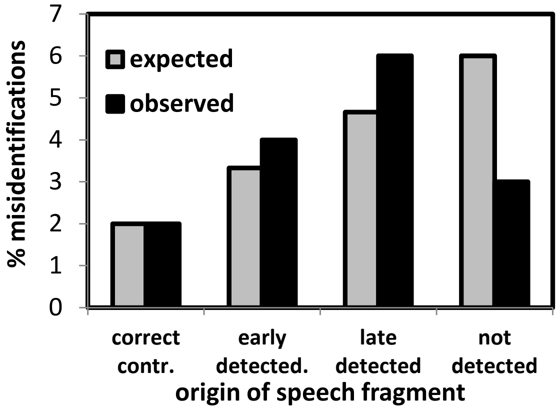

Figure 1: *Relative frequencies (in %) of expected and observed misidentifications of speech fragments for the four origins of speech fragments.*

The expected values were calculated by taking the range in the observed values, running from 2% to 6%, and dividing the other intervals equally, following the predicted order. Obviously, the order of observed values does not follow prediction, mainly because fragments from undetected errors have less, not more misidentifications than those from early and late detected errors. The responses were analyzed by means of mixed-effects logistic regression models with misidentification as a binomial dependent variable [10] and three planned orthogonal contrasts, viz. correct controls versus speech errors, undetected versus detected speech errors, and early versus late detected speech errors. The results of the best fitting model shows that, as predicted, fragments from speech errors have more misidentifications than fragments from correct controls ($p < .0001$), that fragments from undetected errors have less misidentifications than fragments from detected errors ($p < .0085$), and that late detected errors tend to have more misidentifications than early detected errors ($p < .0624$).

Figure 2 presents expected and observed average RTs for the four origins of speech fragments. Only RTs were included for correctly identified speech fragments, and only for those stimuli having not more than 3 misidentifications. Also 29 outliers were removed. There remained 11197 cases, 92% of all responses.

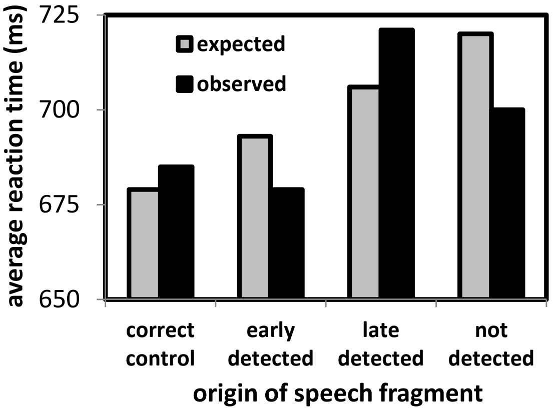

Figure 2: *Expected and observed RTs.*

Predicted values were calculated by starting from the range of observed values, running from 679 ms to 721 ms, and dividing the other intervals equally, following the predicted order. The observed order of average RTs does not follow the predicted order. After RTs were log-transformed to obtain a more normal distribution they were analyzed by means of mixed-effects regression models (LMM) with two crossed random effects, viz. pairs of stimuli (speech error paired with matching control) and listeners [10]. Again there were three planned orthogonal contrasts, viz. correct controls versus speech errors, undetected versus detected speech errors, and early versus late detected speech errors. The best fitting model did not include interactions. The model showed that there was no significant difference between fragments from correct controls and fragments from speech errors, no significant difference between detected and undetected errors, and a significant difference between early and late detected errors. Post hoc comparisons showed that fragments from early detected errors gave significantly *shorter* RTs than correct control fragments ($p < .0182$) and than fragments from undetected errors ($p < .0295$), and that fragments from late detected errors lead to significantly *longer* RTs than fragments from undetected errors.

4. Discussion

Systematic differences in percentages misidentifications and in RTs in a simple identification task with meaningless CV speech fragments can only reflect systematic differences in perceptual clarity of these speech fragments and in particular of the identified consonant segment. The basic assumption underlying the current research, based on the findings in [2] and [4] is that such differences in perceptual clarity of initial consonantal segments are caused by differences in the amount of articulatory ambiguity caused by conflicting articulatory gestures.

The percentages of misidentifications, interpreted in this way, suggest that, as one would expect from the results in [2] and [4], consonant segments resulting from speech errors on average suffer more from articulatory ambiguity than consonant segments in correct controls. They also suggest that consonants in detected speech errors suffer more from articulatory ambiguity than those in undetected speech errors. The effect is particularly strong for late detected speech errors, as if in these latter cases articulatory ambiguity positively correlated with probability that the segmental errors were detected and repaired in self-monitoring. If so, this is not predicted by the perception-based monitor comparing error form with the correct target [3], [7]. The result rather supports a theory of production-based self-monitoring in which the probability of error detection by the monitor increases with the amount of conflict between competing segments [8]. The consonant segments made in early detected speech errors apparently suffer on average less from articulatory ambiguity than segments made in late detected errors. This is in line with our suggestion that in early detected speech errors often speech is initiated before the competing correct target segment is fully activated. This would decrease the probability of articulatory conflict between the two segments and thus reduce the probability of perceptual unclarity caused by articulatory conflict.

It is noteworthy that the above interpretation of the relative numbers of misidentifications is based on only a tiny fraction of all responses. Only 3% of all responses deviated from the auditory transcription made by a single trained phonetician as described in [7] This means that such auditory transcriptions are fairly reliable as descriptions of what people actually perceive when hearing segmental speech errors. But this also means that we should be careful not to draw definitive conclusions from the handful of misidentifications. They are not necessarily representative of how the perceptual clarity of speech segments depends on their origin as correct controls, undetected, early detected and late detected speech errors. More weight should be given to RTs in the identification of those speech fragments that were not misidentified, and did not come from stimuli with more than 3 misidentifications. The reader may recall that the analysis of RTs was carried out on 92% of all responses. Here we assume that differences in RT reflect differences in perceptual unclarity that in turn stem from differences in the amount of articulatory conflict.

An important result is that identification of speech fragments from early detected consonantal errors does not take more but less time than identification of speech fragments from correct controls. This can only mean that on average these segments do not suffer at all from articulatory ambiguity. We interpret this as meaning that (a) speech errors start their life in internal speech as full substitutions of one segment by another, and that (b) often speech is initiated, i.e. activation is passed on to the articulators, immediately after activation of the error segment and before activation of the correct target segment. This is a reasonable hypothesis if we assume that the segment that is most strongly activated also arrives slightly earlier in the speech plan. Early detected speech errors are then precisely the cases where the error segment is more strongly activated than the correct target segment. When, slightly later, the correct target segment is also activated, the monitor can detect the error either on the basis of the amount of conflict between the two segments, or on the basis of perception-based comparison between the two segments. If the error is detected, the speech that is already initiated will be interrupted and a repair will be made. This nicely explains the frequent occurrence of early interruptions both in spontaneous speech errors [5] and in speech errors elicited with the SLIP technique [6], [7].

A second important result is that identification of speech fragments from late detected errors takes considerably more time than identification of speech fragments from early detected and undetected speech errors. We would not expect this on the basis of perception-based comparison between error form and correct target form, because a difference would be most easily detected when the perceptual distance is optimal. Our interpretation of the long reaction times for late detected errors is that these speech errors suffer relatively strongly from articulatory ambiguity and that the amount of articulatory conflict either directly or indirectly has increased the probability of error detection by the monitor. Directly if articulation is monitored for the amount of conflict between competing articulatory gestures, indirectly when the overt speech is monitored auditorily for unclear segments.

5. Conclusions

Speech errors start their life in inner speech as full categorical substitutions. These surface in overt speech in early interruptions where error detection follows speech initiation. Competition between segments surfaces as perceptual unclarity due to articulatory conflict. Probability of error detection increases with amount of conflict.

6. References

[1] B.J. Baars and M.T. Motley. “Spoonerisms: Experimental elicitation of human speech errors”. Journal Supplement Abstract service, Fall 1974. Catalog of selected documents in Psychology 3, pp. 28–47, 1974.

[2] L. Goldstein, M. Pouplier, L. Chen, E. Saltzman and D. Byrd. “Dynamic action units slip in speech production errors”. *Cognition* 103, pp. 386–412, 2007.

[3] W.J.M. Levelt, A. Roelofs and A. S. Meyer. “A theory of lexical access in speech production”. *Behavioral and Brain Sciences* 22, pp. 1–75. 1999.

[4] C.T. McMillan and M. Corley. “Cascading influences on the production of speech: Evidence from articulation". *Cognition* 117, pp. 243–260, 2010.

[5] S.G. Nooteboom. “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*, pp. 167–186, 2005.

[6] S.G. Nooteboom. “Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech”. *Speech Communication* 47, pp. 43–58, 2005.

[7] S.G. Nooteboom and H. Quené. “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, pp. 837–861, 2008.

[8] N. Nozari, G. Dell and M. Schwartz. “Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production”. *Cognitive Psychology* 63, pp. 1–33, 2011.

[9] M. Pouplier and L. Goldstein. “Asymmetries in the perception of speech production”. *Journal of Phonetics* 33, pp. 47–75, 2005.

[10] H. Quené and H. Van den Bergh. “Examples of mixed-effects modeling with random effects and with binomial data”. *Journal of Memory and Language* 59(4), pp. 413–425, 2008.

Proceedings of

DiSS 2013

The 6th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
21–23 August 2013

TMH-QPSR

Volume 54(1)

Edited by
Robert Eklund

Conference website: http://www.diss2013.org
Proceedings also available at: http://roberteklund.info/conferences/diss2013

Cover design by Robert Eklund
Front cover photo by Jens Edlund and Joakim Gustafson
Back cover photos by Robert Eklund

Proceedings of DiSS 2013, The 6th Workshop of Disfluency in Spontaneous Speech
held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 21–23 August 2013
TMH-QPSR volume 54(1)
Editor: Robert Eklund
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISBN 978-91-981276-0-7
eISBN 978-91-981276-1-4
ISSN 1104-5787
ISRN KTH/CSC/TMH--13/01-SE
TRITA TMH 2013:1

Printed by Universitetsservice US-AB, Stockholm, Sweden, 2013