Articulatory Timing of Coproduced Gestures and Its Implications for Models of Speech Production

Marianne Pouplier[1,2] and Susanne Waltl[1]

[1]*Institute of Phonetics and Speech Processing, Ludwig-Maximilians Universität München;*
[2]*Haskins Laboratories*
E-mail: pouplier@phonetik.uni-muenchen.de; susanne@phonetik.uni-muenchen.de

Abstract

Several studies have reported that during phrases with alternating consonants, the constriction gestures for these consonants can come to be produced simultaneously during the same token. Since these coproductions occur in contexts that also elicit segmental substitution errors, the question arises whether they may result from a monitoring and repair process, or whether they arise through the inherent architecture of the speech production system itself. This paper explores the articulatory timing of the coproduced gestures in order to shed light on the underlying process that gives rise to them. Results show that at movement onset the gestures are mostly synchronous, but it is the intended consonant that is released last. For some of the timing values, the intended gesture follows the intruding one with a relatively long lag, supporting a monitoring account. However, similarly long lags are observed in cases in which the temporal order of the two gestures is opposite to the predictions of a monitoring account. The median values in particular are consonant with the view that the activation of two gestures is conditioned by the nature of the speech production process itself and not necessarily the result of a repair process.

1 Introduction

Utterances with alternating consonants – typical environments for the occurrence of speech errors – display an increased amount of articulatory token-to-token variability such that the intended as well as an intruding, errorful gesture can come to be produced during the same token [2, 7, 10]. In earlier work, we have hypothesized that these simultaneous productions of two gestures are due to a dynamic synchronization process during which otherwise alternating gestures come to be produced in a 1:1 frequency mode [2]. Since these coproductions occur in contexts that also elicit segmental substitution errors, and may be perceived as speech errors [11], the question arises whether either speech errors are not necessarily segmental substitutions or whether the coproductions may result from a monitoring and repair process. In the latter view, a segmental substitution may be discovered by a monitoring mechanism and a repair is initiated before the error is fully articulated. Error and repair could then be articulated (near-)simultaneously (cf. [8] for an overview). The co-productions would then only be a secondary consequence of the error as opposed to resulting from the architecture of the speech production system as part of the error itself ([1, 2]). Evidence that some of the coproductions may result from monitoring comes from McMillan [6]. He found for his EPG data that for more than 90% of his tokens, closure for the intended gesture followed closure for the intruding one, as expected in a repair. However, McMillan only relates the timepoints of articulatory closure, yet several studies have shown that the articulatory variability for the intruding gesture spans a continuum of movement amplitudes and we do not necessarily see closure. Using EMMA data, the current paper investigates articulatory timing between the coproduced gestures on the basis of several gestural landmarks and across the full range of variability in articulator height. The aim is to understand whether the coproductions are part of the speech production/error process or arise from monitoring and repair.

2 Data

We analyzed EMMA data for four native speakers of American English. They repeated utterances with alternating onset consonants (e.g., *cop top*) synchronized to a metronome beat for

about 10s per trial; two rates were employed ('fast': 120 beats per minute; 'slow': 80 bpm, set speaker-specifically +/- 4 bpm of target rate). The data recording and processing procedures are detailed in Pouplier [9]; the current data are a subset of the data presented therein. The current data were collapsed across the experimental variables stress, position and vowel which were included in the original dataset.

We will refer to the initial consonant of the word the subject was instructed to pronounce as the *intended* consonant/gesture. The *controlled* articulator refers to the articulator forming the main constriction for a given intended consonant (tongue tip for /t/ and tongue dorsum for /k/). The (by hypothesis) *uncontrolled* articulator refers to measurements of tongue dorsum kinematics during /t/ and tongue tip during /k/. Any labelled kinematic event in the uncontrolled articulator will be referred to as *intruding* gesture.

3 Measurements

On the basis of changes to the tangential velocity profile, the vertical movement time series of each gesture was labelled according to the following temporal landmarks: gesture onset (GONS), plateau achievement (TONS), maximum constriction (MAX), end of plateau (release; TOFFS). These landmarks were defined on the basis of a 20% threshold of the peak tangential velocity of a given trajectory.

The uncontrolled articulator was labelled in addition to the controlled articulator whenever the labelling algorithm identified all gestural landmarks within a given window. Windowsize was always chosen so as to include an intended single repetition of the target phrase. Crucially, this labelling criterion did not rely on an a priori classification of tokens as errorful; the inclusion of any given token in the analysis was solely based on its velocity profile. A total of 1095 tokens was included in the analysis (fast: 632, slow: 463), corresponding to about 54% of all tokens. As measures of interarticulator timing for the coproduced gestures, lags were computed by subtracting the timestamp of the controlled articulator from the timestamp of the uncontrolled articulator for corresponding landmarks. A positive value means that a given landmark of the intruding gesture occurred later than the corresponding landmark in the intended gesture. A negative value indicates that the landmark of the intended gesture occurred later in time.

4 Results

Figure 1 shows the median lag values for the successive landmarks across speakers separately for the two speaking rates. For both speaking rates, there is a trend towards decreasing negative values for the successive gestural landmarks. The two gestures start their movement around the same time with the median being close to zero. With each successive landmark, the gestures drift further apart in time with the intended gesture occurring later in time at TOFFS. This effect is more pronounced for the slow speaking rate compared to the fast speaking rate. In order to understand whether the median values are part of a continuum of values or arise from a bimodal distribution (possibly indicative of different underlying processes), the left graph in Figure 2 shows a histogram of the GONS lag values across all subjects and tokens. The distribution is continuous between +/-250 ms, with only a few tokens outside of that range.

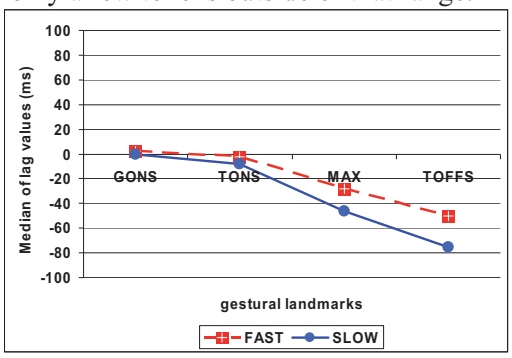

Figure 1: *Median of lag values of the successive gestural landmarks for the two speaking rates*

It is further of interest to consider the range of timing values in more detail. Table 1 gives the minimum and maximum lag values for each landmark for the two speaking rates. The minimum lag values are always negative, indicating that for the extreme cases the intended gesture is always later than the intruding one. The maximum lag values are positive for all landmarks and comparable in magnitude to the negative values. That is, lag values in which the intended gesture is followed by an errorful gesture may be as large as cases in which the intruding gesture precedes the intended one. A tendency for a rate effect is also observable in that the slow rate has mostly longer lags compared to the fast rate. Some lag values are quite large, such as -466 ms or 538 ms for GONS at

the slow rate. Visual inspection of some of these tokens shows that the intruding gesture may be released when the intended gesture begins its path towards the target, that is, the gestures are sequential.

Table 1. *Minimum and maximum lags (ms) for the two speaking rates.*

Rate		GONS	TONS	MAX	TOFFS
fast	Min	-266	-248	-192	-230
	Max	234	196	158	166
slow	Min	-466	-404	-406	-382
	Max	538	484	240	240

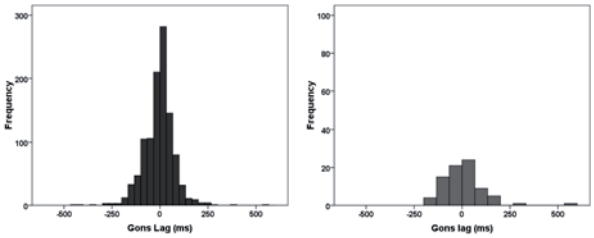

Figure 2: *Histograms of GONS lag values (ms). Left: across all tokens and subjects. Right: including only tokens with highest TDy/TTy values of the uncontrolled articulator.*

Conceivably, the negative and positive lag values of equal magnitude could arise from fact that repairs fall onto the negative side of the continuum, while tokens that present a normal range of coarticulatory variability might predominately fall onto the positive side (or equally to both sides). If there are only few 'true' error and repair tokens in the data, our present analyses might disguise such a pattern. If that were the case, presumably tokens during which the gestural magnitude of the uncontrolled articulator is at the top end of the continuum would show a bias towards negative GONS lag values, because the intended gesture would, as a repair, follow the intruding one. The right hand graph of Figure 2 shows the histograms for GONS lags including only for each subject the 10 /t/ and 10 /k/ tokens with the highest vertical position of the uncontrolled articulator (about 4% of each subject's tokens). Overall, the pattern has not changed compared to the full dataset: There is no indication in the shape of the distribution that different lag values may be conditioned by different underlying processes.

We also examined whether the intended and intruding gestures differ in plateau duration, since this would account for the changing lag values over the course of gestural activation (cf. Table 2). The plateau duration is shorter for the intruding gesture compared to the intended gesture for both speaking rates. A repeated measures ANOVA with the factors Rate (fast, slow) and Gesture (intended, intruding) was significant for the main effects (Rate: $F(1,3)=10.18$, $p=.05$; Gesture: $F(1,3)=48.81$, $p=.006$).

Table 2. *Average plateau duration (ms) and SDs across subjects.*

		plateau duration (ms)	
Rate		intended	intruding
fast	Mean	67.83	42.31
	SD	42.32	40.39
slow	Mean	83.11	70.75
	SD	57.15	66.97

5 Discussion

Overall, the timing between the coproduced gestures varies systematically over the time course of gestural activation and as a function of the intended consonant. The gestures are mostly synchronous at GONS, but the intruding gesture has a shorter plateau duration and it is thus the intended gesture that is released last. When the extreme values of the distributions are taken into account, we see that a comparable range of positive and negative lag values was observed for both temporal orders: intended before intruding and vice versa. The results remain qualitatively unchanged if only the 10 tokens with the highest vertical positions of the intruding gesture are taken into account.

The longer negative lag values speak for a monitoring account such as proposed by Levelt [4]; in his model inner speech is monitored by the comprehension system. In such a model, error correction requires replanning before a repair can be issued and thus a correction might lag an incorrect production by 100 ms or more [8]. Yet not all tokens can easily be accounted for as a repair: the lag values for tokens in which the intended gesture precedes the intruding gesture (i.e., the opposite temporal ordering as predicted by a monitoring account) are as big as the lag values for tokens that can be interpreted as repairs. Also the median values do not speak for the conclusion that these errors can overall be explained on the basis of rapid repair mechanisms. The current data thus underscore the difficulty of separating out 'repair-

tokens' from others. The distribution of the data does not show any evidence for separate underlying mechanisms for different lag values.

The current data show quite discrepant results to the McMillan study [6]. In parts this is surely due to methodological differences, yet the difference in results is substantial: The lag between the two closures exceeded 180 ms for more than 80% of tokens in the McMillan data. Such a time span has been hypothesized to be required to detect an error and initiate a repair [3, 8]. In the present data only 1% percent of tokens had negative TONS lags within that range. McMillan used a word-order reversal task which, in contrast to the present study, did not employ continuous repetition. Yet he identified the same types of gestural coproductions as we did in our data and he argues for coproductions resulting from an interactive speech production process. The impact of the task on articulator kinematics and speech error types thus remains a topic for future research.

The continuum of both negative and positive lag values, as well as the median values of around zero at GONS are in agreement with our interpretation that errors may arise from a gestural synchronization process in which otherwise alternating gestures come to be produced in a 1:1 frequency mode [2]. Yet this account does not explain why the intended gesture is released last. The cascading activation model [1] can provide an account of this phenomenon since it assumes that the strength of activation during phonological planning is directly correlated to articulatory strength. In this context it has also been argued that intended representations will be activated more strongly than their competitors. Plateau duration is a possible correlate of articulatory strength, and in our data the intended gesture has indeed a longer plateau than the intruding one, lending further support to this view. A limitation of the cascading activation model is though that it makes generally no predictions about articulatory timing and it is unclear how to explain the range of timing values observed.

Lastly, it is worth noting that while the articulatory release is generally dominated by the intended gesture, the acoustics are dominated by the tongue dorsum [5]: Overall, tokens with a TD gesture (intended or intruding) are acoustically closer to /k/ than to /t/. This makes it unlikely that the timing patterns reported here are governed by the acoustic output.

Overall, some of the observed lag values are in agreement with a monitoring account. However, the range of values observed and their overall distribution argue for the view that coproductions reflect the inherent nature of the speech production system rather than arising from a secondary monitoring process.

Acknowledgments. Work supported by the Deutsche Forschungsgemeinschaft (PO 1269/1-1) and NIH (R01 DC008780-01).

References

[1] Goldrick, M. & S. Blumstein, Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, 2006. **21**: 649-683.

[2] Goldstein, L., et al., Dynamic action units slip in speech production errors. *Cognition*, 2007. **103**(3): 386-412.

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

[4] Levelt, W., et al., A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 1999. **22**: 1-75.

[5] Marin, S., et al., Acoustic consequences of gestural intrusion errors. *JASA*, 2008. **123**(5): 3329.

[6] McMillan, C., *Articulatory Evidence for Interactivity in Speech Production*. 2008, PhD thesis, Univ of Edinburgh.

[7] Mowrey, R.A. & I.R. MacKay, Phonological primitives: Electromyographic speech error evidence. *JASA*, 1990. **88**(3): 1299-1312.

[8] Postma, A., Detection of errors during speech production: a review of speech monitoring models. *Cognition*, 2000. **77**: 97-131.

[9] Pouplier, M., *Units of phonological encoding: Empirical evidence*. 2003, PhD dissertation, Yale Univ.

[10] Pouplier, M., The role of a coda consonant as error trigger in repetition tasks. *JPhon*, 2008. **36**: 114-140.

[11] Pouplier, M. & L. Goldstein, Asymmetries in the perception of speech production errors. *JPhon*, 2005. **33**: 47-75.