Running head: DYNAMIC ACTION UNITS

Dynamic Action Units Slip in Speech Production Errors

Louis Goldstein[1, 2]

Marianne Pouplier[1, 2, 3, *]

Larissa Chen[1, 2]

Elliot Saltzman[1, 4]

Dani Byrd[1, 5]

[1] Haskins Laboratories, New Haven, Connecticut

[2] Yale University

[3] now at the University of Edinburgh

[4] Boston University

[5] University of Southern California, Los Angeles

* Corresponding author. Marianne Pouplier, LEL, University of Edinburgh, Adam Ferguson Building, 40 George Square, Edinburgh EH8 9LL, UK;

pouplier@haskins.yale.edu

Abstract

In the past, the nature of the compositional units proposed for spoken language has largely diverged from the types of control units pursued in the domains of other skilled motor tasks. A classic source of evidence as to the units structuring speech has been patterns observed in speech errors—"slips of the tongue." The present study reports, for the first time, on kinematic data from tongue and lip movements during speech errors elicited in the laboratory using a repetition task. Our data are consistent with the hypothesis that speech production results from the assembly of dynamically-defined action units—gestures—in a linguistically structured environment. The experimental results support both the presence of gestural units and the dynamical properties of these units and their coordination. This study of speech articulation shows that it is possible to develop a principled account of spoken language within a more general theory of action.

Keywords: speech errors, action units, entrainment, speech production

Introduction

While humans perform skilled acts of motor control in many domains, none have the communicative and information encoding properties of language. This special status of language has no doubt contributed to accounts of speech that make it appear to be quite different from other kinds of coordinated action. While phonological research is built on the insight that words are composed of combinatorial units of information such as features or segments, finding a physical basis for compositional units in spoken language, either in the articulatory actions humans perform when speaking or in the resultant acoustic signal, has proven elusive. There is no obvious discrete division of the speech acoustic signal into compositional units; there is no analogue to the spaces between the letters of written language or a phonetic transcription. However, it has been proposed that when articulatory kinematics are examined within a dynamical systems framework, it becomes possible to identify compositional action units as atomic units of speech production and to understand how general dynamical principles that apply to skilled action generally also shape the activity of speaking (Browman & Goldstein, 1986; Fowler, Rubin, Remez, & Turvey, 1980).

One reason for the difficulty in identifying units in speech production is the difficulty of observing the speech articulators—such as the tongue—in action. Some evidence for action units has been found in studies using experimentally-induced mechanical perturbation of articulators during running speech. These have shown that certain vocal tract articulators cohere systematically in the production of particular information units. For example, the upper lip, lower lip, and jaw appear to work

cooperatively (and compensatorily) to achieve lip closure (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). Similar studies using phase-resetting techniques have demonstrated compensation in the temporal domain (Saltzman, Löfqvist, Kay, Kinsella-Shaw, & Rubin, 1998).

Based in part on such studies (as well traditional phonological investigations into how words in a language are systematically differentiated from one another and how they are modified when they are produced in different contexts), it has been proposed that utterances can be decomposed into sets of dynamically-defined units of constriction action, called gestures (Browman & Goldstein, 1992, 1995). A gesture orchestrates the movements of several articulators (e.g., upper lip, lower lip, jaw) in order to achieve a linguistically-significant goal (e.g., lip closure). Gestures are modeled as point attractors in a task space—where the attractors are, in their simplest form, characterized as critically damped mass-spring systems, and the task space dimensions are defined in terms of constrictions that can be created and released by the independently controllable constricting organs of the vocal tract (Saltzman & Munhall, 1989). The gestures comprising an utterance have activations that wax and wane over time as the corresponding task-space attractors come into and out of existence in the vocal tract. Simultaneously with being compositional units of action, gestures function as combinatoric units , such that words may meaningfully differ from one another in the identities of, and the relative timing or phasing among, the gestures included (Browman

& Goldstein, 1989; Byrd, 1996; Goldstein & Fowler, 2003).[1] In our view, larger units such as segments and syllables are built from the atomic gestural units, and we refer to these as gestural molecules. Gestural molecules can be modeled as temporally coordinated, that is, coupled, assemblies of gestures (Byrd, 1996; Saltzman & Byrd, 1999; Saltzman & Munhall, 1989).

Errors made during speech production have long been used to investigate the compositional units of spoken language. Speech errors are not random distortions; rather they are systematic in their occurrence and distribution (Dell, 1986; Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979). For example, instead of saying the intended phrase "coffee pot," someone might be heard to say something that sounds like "poffee cot" or like "poffee pot" but not like "cottee poff"—this is because consonants are likely to interact in errors if they share the same word or syllable position (cf. Meyer, 1992 for an overview). Since errors seem to obey the laws of phonology, units that participate in errors are considered to be significant cognitive units of planning in normal, error-free speech, and results from speech error research have thus played a pivotal role in shaping

---

[1] The lexical items bad and mad, for example, minimally differ from each other by the presence or absence of a word-initial velum lowering gesture. The relative phasing of a velum-lowering gesture relative to an oral stop gesture, for instance, determines the difference between the words ban and band.

the architecture of speech production models (e.g., Dell, 1986; Levelt, 1989; Shattuck-Hufnagel, 1979).[2]

For decades, linguists and psychologists have kept transcription records of the speech errors they hear—or overhear—in natural settings (Fromkin, 1971; Meringer & Mayer, 1895; Shattuck-Hufnagel & Klatt, 1979), and experimental paradigms have been developed to elicit speech errors in the laboratory (see e.g., work by Motley & Baars, 1976; Stemberger, 1991). Error rate in normal speakers is very low, about 0.1-0.2% (Garnham, Shillock, Brown, Mill, & Cutler, 1981). Since data acquisition time in the laboratory is constrained, many studies of phonological speech errors have used a repetition task, similar to tongue twisters, so as to elicit a meaningful number of errors per subject (e.g., Dell, Reed, Adams, & Meyer, 2000; Goldrick & Blumstein, in press; Wilshire, 1998). Shattuck-Hufnagel (1983) obtained comparable results for laboratory elicited errors using a repetition task and the analysis of naturally occurring errors. Based on these transcription records of naturally occurring and laboratory elicited errors, several generalizations have been made about the nature of errors that have implications for the compositional units involved in speech production. The basic observation is that the most common sound error is substitution of one segment-sized unit for another (Dell, 1986;

---

[2] More recently, speech production models have increasingly drawn on reaction time data, for instance from picture naming paradigms (e.g., Levelt, Roelofs, & Meyer, 1999). However, speech errors continue to be used as key evidence for the psychological reality of linguistic units in the lexicon .

Fromkin, 1971, 1973; Meyer, 1992; Shattuck-Hufnagel, 1979, 1983).[3] The second important observation about errors is that they produce an utterance that is grammatically well-formed in the language; that is, they correspond to actual or possible words (Shattuck-Hufnagel, 1983; Shattuck-Hufnagel & Klatt, 1979, but see e.g., Laver, 1979) This has been interpreted as evidence for a production frame that allows only units of a particular type in a given slot (but see, for example, Dell, Juliano, & Govindjee, 1993 for an alternative account). For instance, vowels only substitute for vowels, and consonants only for consonants. Crucially, low-level properties such as aspiration or vowel tenseness of a serially misordered segment correspond to the new environment: That is, only its serial order distinguishes a normally from an abnormally produced sound, indicating that errors arise from a temporally inappropriate selection of abstract, symbolic phonological units.

Whether errors are collected in "natural settings" or in the laboratory, the basic research tool for speech error studies has been one and the same: phonetic transcription. Yet there are several critical reasons why transcription studies of speech errors are an incomplete source of evidence for the nature of speech production units (see Boucher, 1994; Cutler, 1981; Ferber, 1991). Such studies rely inherently on a written record of how an utterance is perceived, usually recorded as a transcribed sequence of segments.

---

[3] Also, errors in which segments are inserted are reported to be more common than errors in which segments are deleted (Stemberger, 1982, 1991; Stemberger & Treiman, 1986).

They do not, and in fact cannot, provide a record of the articulatory events that produced the acoustic signal presented to the transcriber. In particular there is no way to record the presence of a constriction gesture when it is not isomorphic with some transcribed segment, that is, subsegmental. Relatedly, a partial but "incomplete" articulatory constriction movement may fail to be recorded since partial or gradient errors may have very little effect on the acoustic signal, making them imperceptible or close to imperceptible to the listener (and even if some anomaly is heard, transcription systems provide no obvious way to record it). In fact, several studies that have examined the acoustic signal during speech errors quantitatively have suggested that such gradient errors do occur (Frisch & Wright, 2002; Goldrick & Blumstein, in press; Laver ,1979). These suggestions are further supported by an electromyographic study of speech errors that found aberrant activation of muscles such that "parts of motor patterns of ... two items are produced at the same time" and over a range of magnitudes (Mowrey & MacKay, 1990; cf. also Boucher, 1994).[4] In fact such patterns were reported to be frequent rather than unusual. However in this EMG study, the precise nature of the articulatory events that corresponded to the experimental observations remained

---

[4] Gestural units, as coordinative structures, call on several muscles and/or articulators. Errors at the gestural level are thus very different from errors at the level of individual muscles, claimed by Mowrey and MacKay (1990). However, since they measured single muscles, there is no way to know if the errors they observed are in fact occurring at the level of gestures or muscles.

unknown. In particular, since only a single muscle site was recorded, it is possible that the gradient error activity could have been counteracted by activity of other muscle(s) that prevented actual errorful movement of the tongue. Thus, the apparent gradient error could simply reflect non-errorful variation in the levels of co-contracting muscles. Also, acoustic investigations of speech errors do not allow for an unambiguous interpretation of the articulatory events that have rendered a particular acoustic outcome, due to the complex relationship between acoustics and articulations (Atal, Chang, Mathews, & Tukey, 1978; Chen, 2003).

Since speech errors provide a crucial source of evidence for the nature of units in speech production, the inadequacy of past experimental work for supplying information about action units should be remedied. Kinematic data, that is, articulatory movement observation, provides an appropriate record of speech production errors and, by extension, informs more adequately as to the existence and nature of speech action units. We present such data for the first time. To foreshadow, the kinematic data provide evidence for a kind of error (partial gestural intrusions) that cannot exclusively be explained as result of simple substitution and/or exchange of segmental units. Such errors are predicted by the view that speech production is based on action units that are coupled to one another. In these speech errors, constriction actions appear to be activated, to a greater or lesser degree, at incorrect temporal locations in the production of an intended word.

If the atomic units of speech production are, as we suggest, best described in terms of dynamical action units or gestures, the properties of errors as revealed in the

kinematic data should be consistent with the hypothesized dynamic nature of these underlying units. First, the errors should be obviously interpretable in terms of linguistically-significant vocal tract constrictions. A temporally mislocated gesture will not result in random movement; that is, errorful actions will preserve their properties as linguistically significant units of speech production. As such, errorful activity will be directly related to gestural structure. Second, errors should be sensitive to the larger temporal and dynamic context, such as speech rate. Gestural coupling may become weaker under fast speaking rates; (Kelso, Scholz, & Schöner, 1986; Krakow, 1999; van Lieshout, Hulstijn, Alfonso, & Peters, 1997), thus selectively destabilizing certain intended patterns and potentially leading to more errors.[5] Third, we should find errors that involve individual constriction gestures, as well as those that involve larger gestural assemblies—in particular entire segments. If gestures are basic units of speech production, we expect to see them participate individually in errors. Since gestures are hypothesized to be coupled to one another with different degrees of cohesion (Browman & Goldstein, 2000), we should likewise be able observe larger structures to participate in errors. Fourth, speech has an inherently rhythmical basis, as evidenced for example by its metrical structure and its general alternation of vowels and consonants and, as such,

---

[5] Rate effects have been observed in the transcription-based literature (e.g., Kupin, 1979; MacKay, 1971) and have been modelled on the basis of decay rates of previously activated nodes in the lexical network being too slow at high speaking rates (e.g., Dell, 1986).

aspects of speech organization can be illuminated by models of oscillatory dynamical systems. These types of models have been studied in other movement-related domains (e.g., finger tapping), and it has been shown that such systems typically display preferred modes of coordination and characteristic transition stages to these preferred modes. We would expect these dynamical concepts to provide important explanatory power in understanding the emergence of speech errors. In the experimental speech error data that are reported below, we find evidence supporting all four of the above expectations. These findings support the hypothesis that gestural action units are atomic compositional units in spoken language and that a dynamical model of these units can, contrary to previous claims (e.g., Levelt et al., 1999), account for properties of speech errors.

Experiment 1: Looking Inside the Mouth During Errors

Method

A repetition task was employed to elicit speech errors while collecting articulatory movement data with an electromagnetic mid-sagittal articulometer (EMMA, Perkell, Cohen et al., 1992). The apparatus allows the tracking of individual fleshpoints by means of small transducer coils attached to various points on the subject's vocal tract in the midsagittal plane. Transducers were placed on the tongue tip, tongue body, tongue dorsum, upper lip, lower lip, jaw, maxilla and nose ridge. Articulatory data were sampled at 500 Hz and smoothed with a low pass filter of 15Hz. Acoustic data was sampled at 48kHz for one subject and 20kHz for all other subjects with a Sennheiser short gun microphone. Speech rate was controlled by means of a metronome; the metronome voltage was sampled at 1000 Hz simultaneously with the speech signal and the EMMA

data. Stimuli were presented on a computer screen positioned about 1 m away from the subject; the metronome (with a light display) was placed on top of the computer screen.

Stimuli and procedure.

Subjects were instructed to repeat two-word phrases with alternating syllable onset consonants—such as in *cop top* or *tip kip*—at three speech rates, which were elicited by means of the metronome. The rates ranged between 76 and 120 beats per minute and were incremented at 20 bpm intervals, allowing for subject-specific adjustments within a 4 bpm range; typical rates were: fast—120 beats per minute/500 ms per repetition, medium—104 bpm/578 ms per repetition, and slow—80 bpm/750 ms per repetition. Trial duration varied with rate: 10 s for the fast rate, 13 s for mid, and 15 s for the slow condition. Other variables manipulated were stress (initial vs. final), word ordering (*cop top*, *top cop*), and vowel (*cop top*, *kip tip*); variables were fully crossed. Tokens intended to elicit errors, that is, those with alternating initial consonants (such as *cop top*), were randomly intermixed with control stimuli having nonalternating consonants (such as *cop cop*, *top top*). The complete set of alternating consonant stimulus phrases were: *cop top, top cop, tip kip, kip tip*. The complete set of control phrases were: *cop cop, top top, kip kip, tip tip*.

The rate conditions were blocked; the order of these rate blocks was changed for each subject. Within each rate block, stress conditions were blocked and subjects were alerted by the experimenters as the stress condition changed. Stress was indicated in the stimulus display by capitalizing the stressed word. Vowel and phrase position conditions appeared in the same random order across the different stress and rate blocks.

Participants.

The data presented are from seven native speakers of American English with no reported speech or hearing deficits. All were all naive as to the purposes of the experiment. One subject performed in only two of the rate conditions (fast & slow), and three performed in only the fast rate condition.

Results

The control utterances (i.e., those without alternating consonants) show, as expected for these error-free productions, a tongue tip raising gesture towards the alveolar ridge to form a constriction during /t/, while the tongue dorsum shows no activity for this consonant. The complementary situation exists during /k/ in control tokens: the tongue dorsum shows a raising gesture (to form a constriction with the soft palate) but the tongue tip does not. There is small movement of the tongue tip during /k/ which is assumed to be passive consequence of the motion of the more massive and biomechanically coupled tongue rear. The times of vertical position maxima recorded for the tongue tip and tongue dorsum transducer coil, respectively, were used as the temporal location of constriction achievement for the tongue tip and tongue dorsum gestures, respectively. The movement time functions obtained through the EMMA system were thus evaluated by finding and marking the relevant vertical position maxima of the transducer coils using software algorithms developed at Haskins Laboratories. If the labelling algorithm did not find a maximum at a point in time relevant for the analysis, its value was measured at the time of a maximum in another signal which the algorithm had identified. For instance, if there was no vertical position maximum for tongue dorsum

during /t/ (since the tongue dorsum is not expected to rise during /t/, only the tongue tip will exhibit substantial movement), tongue dorsum was measured at the time of the tongue tip maximum. The vertical positions themselves at the maxima were used as a measure of constriction: the higher, the more constricted.

One trial (14 successive repetitions) from the error elicitation condition of *cop top* is shown in Figure 1 with the repetitions overlaid.

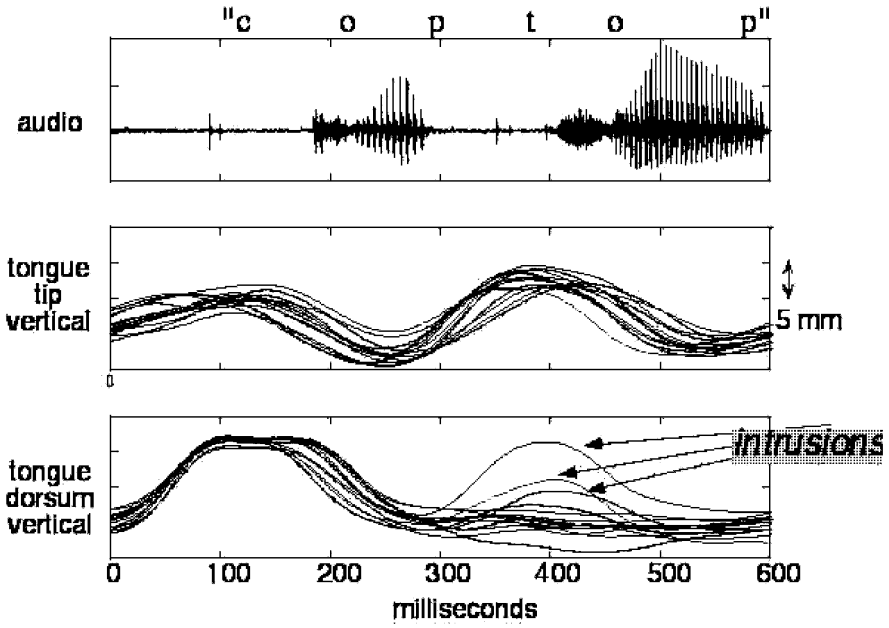

Figure 1. Overlay of 14 successive repetitions of the phrase *cop top* by a single speaker (JP). Top panel shows audio signal from a single repetition. Middle panel shows overlaid time functions of the vertical position of the receiver placed on the tongue tip. Bottom panel shows overlaid time functions of the vertical position of the receiver placed on the tongue dorsum. Line-up is point of zero velocity of vertical tongue dorsum movement.

If this phrase had been performed without errors, like the control utterances, no substantial tongue dorsum movements would be observed during the /t/ in *top*. However, errors did occur in that during some repetitions of *cop top* an "extra copy" of a tongue dorsum constriction appears to be activated in the pre-vocalic position and intrudes during the tongue tip gesture for /t/, which is still produced in that same pre-vocalic position. We will refer to this type of error—that is, the addition of a gesture not produced at that temporal location in a normal, non-errorful production—as a "gestural intrusion error." Intrusion errors of the tongue tip are also seen during /k/. Much less frequently we observed "gestural reduction errors," which we define as an inappropriate reduction in the magnitude of an intended articulatory movement, as for instance a smaller tongue dorsum raising during /k/. Both intrusion and reduction errors were observed during /t/ and /k/ in the alternating-consonant error elicitation trials for all seven subjects.

An error metric was designed based on the statistical properties of the control utterances (mean & standard deviation): any token further than 2 *SD* from its appropriate control mean is considered to be an error. Constriction errors in the form of intrusion or reduction were observed to lie along a continuum of constriction formation, varying from zero to full constrictions. An intruding gesture can be activated partially; likewise, a target gesture can be partially reduced. In order to partition the continuum of constriction magnitude values into "full" or "partial" types, tokens that are both 2 *SD* from the control mean for the intended target consonant as well as 2 *SD* from the other (alternating) consonant were classified as partial, whereas tokens that were 2 *SD* from the control

mean for the intended target but less than 2 *SD* from the control mean for the "unintended" alternating consonant were classified as full. Upon inspection of the variability displayed in the control condition as a function of the experimental factors, error thresholds were calculated separately for each rate and vowel condition since collapsing these factors greatly increased the variability of the measured movement amplitude in the control condition; the values of the control utterances were however collapsed over stress and phrase position conditions.

As would be expected, an alternating utterance has slightly different (co-)articulatory properties than a non-alternating control utterance. This poses the problem that for some trials the overall distribution of constriction magnitude is too different from the non-alternating control distribution for reliably determining what constitutes an error. A threshold of 75% was thus used in order to constrain the power of the error evaluation metric: If the error metric resulted in an error rate of 75% or higher for a given trial, that trial was excluded from further analysis. Statistical procedures were corrected for unequal token numbers.[6]

---

[6] Note that our error metric does not depend on there being no errors at all during the controls. However, errors are assumed to occur only rarely in the non-alternating conditions. If indeed a control trial were to contain many errors, the variability would rise as the result of an overlapping distribution for /t/ and /k/, and the given rate-vowel condition would have to be excluded from analysis.

The most commonly reported error type in studies using transcription data is single segment substitution. A true "substitution" of a compositional unit would be expected to show itself in our experiment as a full reduction of the intended gesture accompanied by a full intrusion of an erroneous gesture. However, the kinematic data show instead, for the conditions of this experiment, that the dominant error production pattern is a quite different one (for studies using related elicitation methods cf. among others Dell et al., 2000; Shattuck-Hufnagel, 1983; Wilshire, 1998). For all seven subjects, there was a strong bias for gestures to be added but not reduced in errors (i.e., a bias favoring intrusion errors over reduction errors). The summary of results can be found in Figure 2. Bars show the percentage of tokens that fell into one of the following three categories: intrusion (either full or partial), reduction (either full or partial), and substitution (simultaneous intrusion and reduction). Note that for two subjects (AB, LK), reduction occurred only when accompanied by intrusion. Overall error rates pooled over all subjects (out of 2648 tokens) are: intrusions: 28.2%, reductions: 3.3%, substitutions: 4.3%.

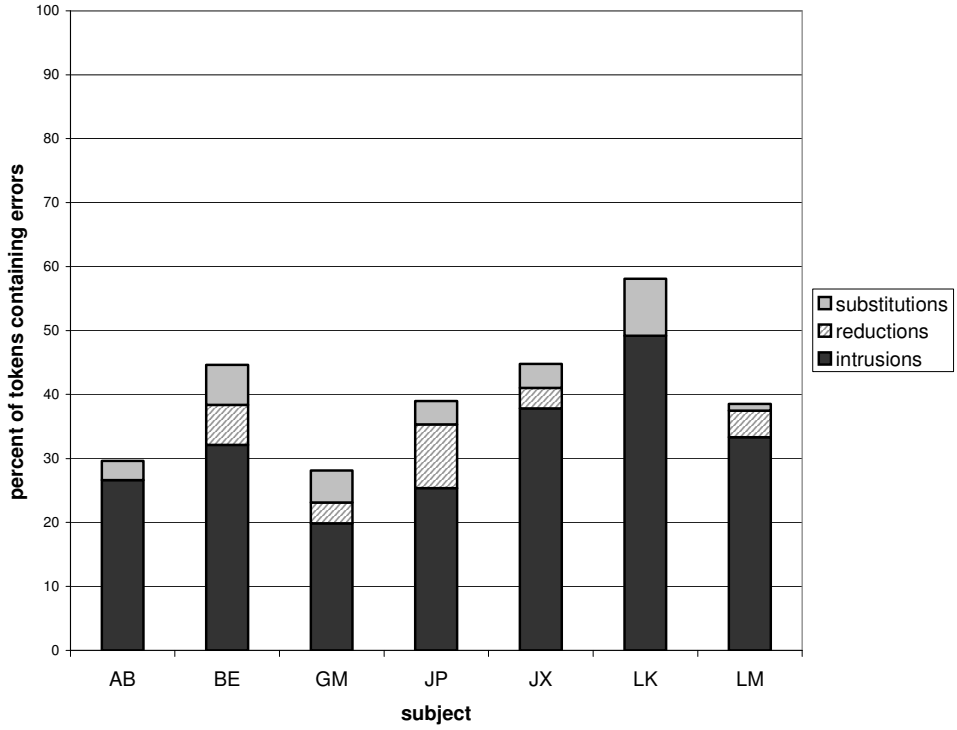

Figure 2. Distribution of error types for all seven subjects.

Of all errors, we find that across subjects 53.8% are partial intrusion errors, while 15.97% are partial reduction errors. 27.3% of all errors fall into the full intrusion category, while 2.9% are full reductions. To investigate whether the intrusion bias was statistically significant for both partial and full errors, an ANOVA with repeated measure on both factors was conducted with the factors Error Type (intrusion vs. reduction) and Error Magnitude Category (full vs. partial). An alpha level of .01 is used for all statistical tests. Both main effects are significant (Error Magnitude Category: $F(1,6) = 14.12$, $\underline{p} = .009$; Error Type: $F(1,6) = 99.04$, $p < .0001$), due to the overall dominance of partial over full errors and of intrusion over reduction errors. The interaction ($F(1,6) = 6.61$, $p = .042$) is not significant, meaning that the intrusion bias holds for full as well as for partial

errors. The weak trend arises from the intrusion bias being slightly stronger at the partial compared to the full level. Overall, partial and full errors follow the same pattern; that is, the systematic dominance of intrusion over reduction is not limited to either the partial or full error magnitude category.

Errors that do not conform to the "add-don't-delete" pattern are rare. A full reduction of a target gesture without an accompanying intrusion error occurs in only 3 out of 946 tokens with errors, while full reduction with simultaneous intrusion of either full or partial magnitude occurs only sporadically (16 out of 946 tokens with errors). Independence of intrusion and reduction was rejected (chi-squared = 57.2, $p < .01$).[7] This bias for intrusion over deletion can also be seen in corpora of transcribed errors in the special case for which multiple gestures can be accommodated by the transcription system, namely when consonant clusters are involved, that is, cases in which two consonants are produced sequentially (Stemberger, 1991; Stemberger & Treiman, 1986).

In traditional speech error research, the term "substitution" denotes the complete, holistic replacement of one segment by another segment. The theoretical framework that has motivated the current study predicts that what has traditionally been thought of as a holistic substitution or replacement is in fact a partial process. Within the terminology employed here, a full substitution corresponds to the simultaneous occurrence of a full intrusion and a full reduction error on the same token. This type of error occurs

---

[7] Contingency table data for total number of errors: Intrusion only—746, reduction only—87, both—113, neither—1702.

regularly—on 4.3% of all tokens—but with lesser frequency than intrusion errors alone. Even in cases when an intrusion and reduction error co-occur on the same token, one or both errorful gestural activations can be of full or partial magnitude.

The role of time and rate.

For the test (alternating) stimuli, the first repetition or two may be easy, but after several, the difficulty often becomes overwhelming. Figure 3 shows the error rate (average normalized error frequency per token)[8] as a function of repetition number pooled over all subjects. Note that for the fast metronome rate, there is a steep increase in error rate from repetition 1 (average error rate of .09) to repetition 7 (average error rate of .32). For the slow metronome rate, error rates are lower, and the buildup less dramatic. The medium metronome rate is intermediate in error rate.

[8] Due to the large variability in the number of tokens obtained in each condition, all error numbers in the following statistics are normalized relative to the number of measured vertical articulator positions per token available for a given subject for the variable under consideration. For instance, when comparing the number of errors for initial versus final stress conditions, for each subject, the number of gestural errors in each stress condition is divided by the number of measurement points (two for each token: tongue tip, tongue dorsum) available for each stress condition for each subject.

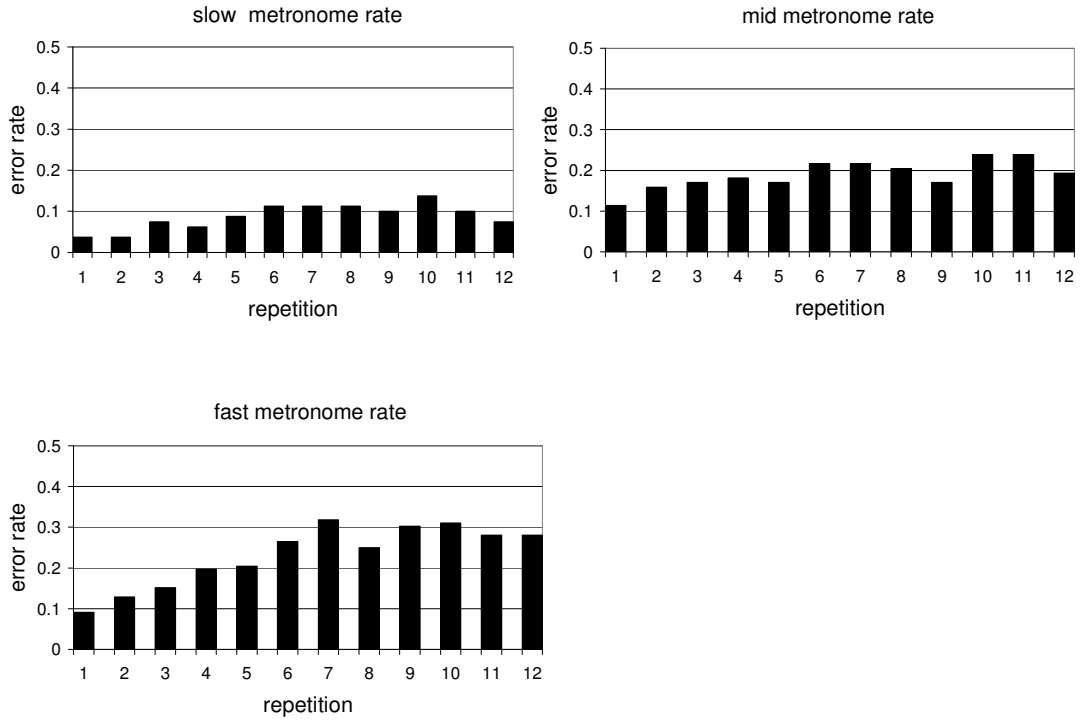

Figure 3. Error rate pooled across the seven subjects as a function of repetition number for each metronome rate.

An ANOVA was conducted to determine whether more errors occurred at faster rates and whether errors built up over the course of a trial. The first 12 tokens of each trial were included. Trials for which there were fewer than 12 repetitions were excluded from the analysis (five trials across subjects). Error numbers were collapsed across the seven subjects and normalized for total number of measurements. Further, the tokens included were marked for whether they occurred at the beginning (tokens 1-4), the middle (tokens 5-8) or the end (tokens 9-12) of the trial. The ANOVA had the factors Rate (fast vs. mid vs. slow) and Trial Part (beginning vs. middle vs. end); both main effects are significant (Rate: $F(2, 36) = 73.3$, $p < .0001$; Trial Part: $F(2,36) = 28.04$, $p <$

.0001); the interaction approaches significance ($F(4,36) = 3.9$, $p = .013$). This means that the number of errors differs significantly with rate as well as with time. A posthoc test (Ryan-Einot-Gabriel-Welsch Multiple Range Test, henceforth REGWQ) on the main effects shows that all three rate conditions differ significantly from each other. For the trial part effect, the middle and end section of the trial are not significantly distinct from each other, but there are significantly less errors in the beginning part compared to either the middle or the end section of a trial.

Qualitatively, the error patterns are comparable across rates, showing a dominance of intrusion without concomitant reduction, as can be seen in Table 1.

Table 1

Error Occurrence and Error Rate Broken Down by Error Type and Speech Rate.

	error type		
metronome rate	intrusion	reduction	both (substitution)
slow	99 (.13)	20 (.03)	17 (.02)
mid	241 (.28)	28 (.03)	41 (.05)
fast	396 (.37)	39 (.04)	55 (.05)

While clearly the fast rate elicits most errors of all types, qualitatively, error distributions are similar in that all rates exhibit a strong intrusion bias, with reduction errors and substitution errors occurring substantially less frequently.

Vowel, phrase position and stress.

A two-tailed matched samples t-test was conducted to compare error numbers on the initial consonants for the two vowel conditions (*top cop* vs. *tip kip*). Average error rate across subjects is .27 (*SD* = .08) for /ɪ/, and .16 (*SD* = .05) for /ɑ/. The difference is significant at $p = .004$ with $t(6) = 4.469$, indicating that there are significantly more errors in the /ɪ/ condition than in the /ɑ/ condition. This may be due to the fact that the SDs of the controls have a tendency to be smaller for the /ɪ/ condition than for the /ɑ/ condition. Thus, for those cases, smaller magnitude deviations (from the control means) will exceed the error threshold for alternating /ɪ/ trials than alternating /ɑ/ trials. In turn, this difference in SD could reflect the fact that the tongue shape produced by the constriction for /ɪ/ is more compatible with the constrictions for /k/ and /t/ than the tongue shape produced by the constriction for /ɑ/. This point will have to be addressed in more detail future research. [9]

[9] An anonymous reviewer suggests that the vowel effect can be accounted for in terms of a resetting effect conditioned by the greater spatial distinction between the constriction locations for the low vowel and either stop consonant compared to that between the high vowel and each consonants. That is, the low vowel may reestablish a tongue position from which movement to either consonant is less error-prone.

The two remaining experimental variables are phrase position (initial vs. final) and stress (initial vs. final). Table 2 shows the mean normalized error frequency with standard deviations in parentheses.[10]

Table 2

Mean Normalized Error Frequency (and SD) for Number of Errors Depending on Phrase Position and Stress.

	phrase position	
stress	phrase initial	phrase final
final	.25 (.05)	.15 (.08)
initial	.19 (.08)	.15 (.07)

An ANOVA with repeated measures on both factors shows that neither stress nor position have a significant effect on error numbers (Stress: $F(1,5) < 1$; Position: $F(1,5) = 6.6$, $p = .051$; interaction $F(1,5) = 1.4$, $p = 2.91$).

Discussion

The common errors observed in our experiment *cannot* be interpreted as arising solely from a process in which one phonological segment substitutes for another. Our data makes clear that, for the experimental conditions employed, an error usually

[10] Data from only six subjects are included in this analysis, as for one subject (LK), for initial phrase position, only the final-stress trials were recorded.

involves the concurrent production of more than one gesture—one appropriate and one intruding.[11] Also, our data (as Mowrey and MacKay (1990) also noted, based on their EMG data) contradict the common claim that errors yield phonologically (i.e., grammatically) well-formed sequences. Concurrent production of tongue tip (/t/) and tongue body (/k/) constriction gestures at the beginning of a word is not licensed by English phonology. Although unit substitution cannot account for the most frequent type of error we observed, it is, however, possible that gestural intrusion errors like the ones we report here underlie the common error types reported in the transcription literature. In separate work (Pouplier & Goldstein, 2005), it is demonstrated that intrusions that are large in magnitude may be perceived by listeners as segmental substitutions (thus accounting for the frequency of transcribed substitution errors), but those that are of smaller magnitude are often imperceptible, meaning that such tokens are perceived as well-formed and error-free *despite* the fact that articulatorily they are clearly not (thus accounting for the apparent well-formedness).

Why do gestural intrusion errors occur?

The strong bias in favor of intrusion over reduction appears at first blush to present a bit of a puzzle. One commonly held belief about speech production (see e.g.,

---

[11] In fact, Experiment 1 is not designed to distinguish whether errors involve individual gestures or gestural molecules (i.e., the segment sized unit for [t] or [k] that includes its laryngeal abduction gesture along with the oral constriction gesture). Experiment 2 will address that question.

Boersma, 1998; Lindblom, 1983) is that speakers attempt to minimize "articulatory effort," however vaguely defined, while speaking. Oddly enough, in production errors we seem to observe quite the opposite: more gestures than required are being produced (cf. also Pouplier, 2003a). Even more puzzling, perhaps, is the fact that these multiple gestures are coproduced in a way that does not correspond to any of the gestural patterns that occur in the speaker's language (for example, the simultaneous coproduction of a /t/-like tongue tip and /k/-like tongue dorsum gesture is not part of any dialect of English).

If action units and their coordination are defined in the language of dynamical systems, a possible explanation for the general bias towards intrusion in errors lies in the behavior of coupled oscillators ("periodic attractors" or "limit cycles"). Activations of individual constriction gestures are orchestrated according to underlying intergestural oscillatory planning dynamics. (See Goldstein et. al., in press, and Saltzman, Nam, Goldstein, & Byrd, to appear for a more extended technical account of the how coupled oscillator dynamics can provide the basis for an utterance's planned pattern of intergestural coordination). For example, in an utterance like *top top*, the tongue tip (/t/) and lip (/p/) gestures exhibit 1:1 frequency locking—one cycle of the tongue tip constrictor is associated with one cycle of the lip constrictor, and the two oscillations are in an anti-phase relation. 1:1 frequency-locked modes are known to be the most stable of the set of possible *m:n* frequency lockings (Haken, Peper, Beek, & Daffertshofer, 1996), and under certain conditions (for example, increased rate), modes with more complex frequency ratios will exhibit transitions to simpler, more stable ratios (Haken et al., 1996; Peper, Beek, & van Wieringen, 1995). In multistable regions of parameter space in which

several frequency-lockings are possible, such transitions can be viewed as resulting from stochastic fluctuations (in frequency or amplitude of the oscillators) that allow the system's behavior to be captured by a more stable ratio, one with a larger basin of attraction. In an utterance like *cop top*, the lip constrictor is in a 2:1 frequency relation with the tongue tip constrictor: two cycles of lip oscillation (/p/) are completed for every one cycle of tongue tip constriction (/t/). Likewise, the relation between the lip constrictor (/p/) and the tongue dorsum (/k/) constrictor is 2:1.[12] The (partial) gestural intrusion errors can then be viewed as the system being captured by the more stable 1:1 mode of frequency locking. In this stable mode, every "word" would have tongue tip *and* a tongue dorsum gesture at the beginning, and a lip gesture at the end. The fact that 1:1 frequency locking appears to be achieved through "extra" cycles of the tongue tip and/or tongue dorsum oscillators, rather than through eliminating a cycle of the lip oscillator is consistent with the results of bimanual tapping experiments that have shown a dominance of the higher-frequency oscillator in mode locking transitions (Peper & Beek, 1999;

[12] It is also possible that the relevant 2:1 coupling relation is between initial consonant oscillators (tongue tip /t/, tongue dorsum, /k/) and the vowel oscillator (tongue root for /ɑ/).

Peper et al., 1995). An example of such a transition to the more stable 1:1 mode can be seen in a token of *cop top* in Figure 4.[13]

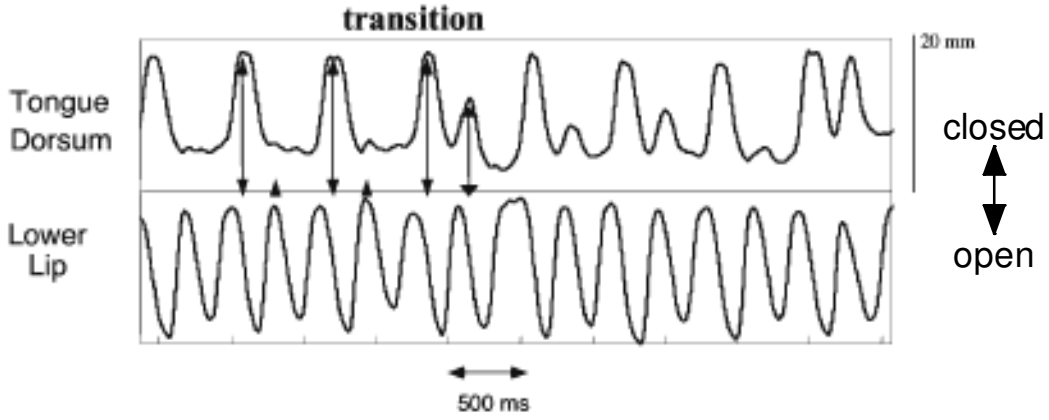

Figure 4. One trial of repeated *cop top* showing the transition of 2:1 coupling of the tongue dorsum and lower lip to 1:1 coupling.

Thus, while these speech errors do not result in an energetically minimal production in terms of the number of gestures produced, the change in gestural pattern

---

[13] Following the dorsum intrusion it can be seen that the lower lip movement is also slightly irregular in that it is spatially and temporally extended. Since the initial consonant gestures are part of the coupling structure for the entire word or utterance, fluctuations can be expected to some extent to increase variability in neighboring gestures as well. This remains a topic for future research.

appears to be governed by dynamical stability principles, according to which gestural patterning is captured by the most intrinsically stable (i.e., 1:1) structure.

This view of the emergence of intrusion errors can also account for the reduction errors and their apparent dependence on intrusions. When gestures are assembled into larger forms (e.g., words), they exhibit certain regular patterns of temporal coordination that develop in the course of learning to speak (indeed, such patterns differ somewhat from language to language). These patterns can be understood as stable *lexical* modes of gestural coordination (phasing or timing). While intrusion errors produce *intrinsically* stable 1:1 frequency locks between the tongue constrictions and the lip constrictions, they have the parallel consequence of pushing the production system to a less stable region of the learned (i.e., lexical) coordination mode landscape. This is because having synchronous tongue tip and tongue dorsum closure gestures at the beginning of a word is not a stable pattern in English; that is, there are no words in the lexicon that begin with this gestural patterning. For this reason, the concomitant reduction of an intended gesture when another gesture has intruded can be understood as shift toward a lexically stable learned coordination mode, one in which there are not two synchronous closure gestures. In general then, speech production errors can be seen as resulting from the interplay of intrinsically stable modes of frequency locking and learned lexical coordination modes. Examination of the error probabilities is consistent with this interplay analysis. Reduction is much more likely on a token that also has an intrusion error than on one which does not. As reported above, independence of intrusion and reduction was rejected.

To underscore this interplay between the two stable modes of coordination, the distribution of substitutions with repetition number within a trial should be considered. Since substitutions are hypothesized to arise from the interplay of a lexically stable attractor with an intrinsically stable attractor, it is predicted that substitutions should build up over time in a similar fashion to what is observed for errors overall. If substitutions only occurred at the beginning of a trial while the co-production of two gestures occurred predominantly in later parts of a trial, this would not fit comfortably in the present account and would point instead to a different process underlying these substitutions. The following figure displays the occurrences of all substitution errors for Experiment 1 across subjects by repetition number within a trial. All repetitions across subjects have been included; the frequency of substitutions per repetition has been adjusted for the different number of repetitions per trial across subjects. While a few substitutions could be observed to occur on the first repetitions of a trial, overall, the rate of substitutions strongly increased over time. This corroborates the hypothesis that substitutions can arise from an interplay of two different stability modes.

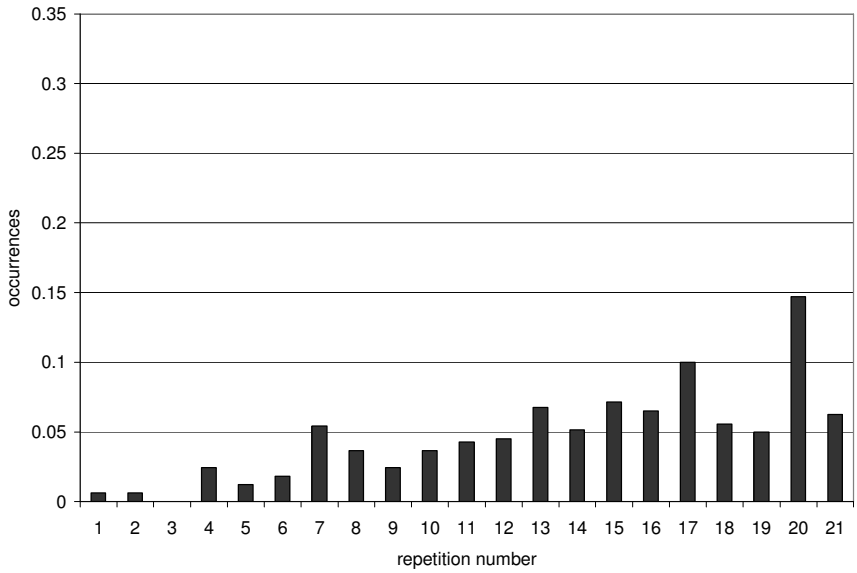

Figure 5. Substitution rate per trial repetition number across subjects.

Finally, we can ask if the fluctuation and capture account can be extended to spontaneous errors in conversational speech, which does not at first glance appear to involve any oscillations.[14] One prediction made by our proposed account is that errors on initial consonants should be more frequent when they involve words that share the same final consonants, as the final consonants in such cases create a higher frequency oscillator that serves to entrain the alternating initial consonants to it, capturing the system in a 1:1 mode. This prediction is borne out both in spontaneous errors and in errors elicited in the

[14] It is worth noting in this regard that mechanical perturbations delivered to the lower lip during speaking have been found to induce compensatory patterns in *both* repetitive and non-repetitive utterances that appear to be governed by the *same* underlying dynamical processes (Saltzman et al., 1998).

laboratory (Dell, 1984). It is also known from transcribed speech error corpora that if segments appear twice in close vicinity, this can increase the likelihood of a speech error (MacKay, 1971)—again consistent with the capturing account. To the extent that laboratory elicitation methods of the kind employed here replicate the processes of normal speech, these data are consistent with a gestural account of speech planning.

The role of time.

The dramatic increase in error rate over repetition can also be modeled as a phenomenon associated with the coupling functions in a set of nonlinear oscillators. As originally formulated by Haken, Kelso, and Bunz (1985) for 1:1 frequency-locked bimanual rhythmic movements, and later generalized to *m:n* frequency-lockings (Haken et al., 1996), inter-pattern phase transitions or bifurcations are induced by nonlinear coupling functions that are rate-dependent. At higher oscillation frequencies, these couplings favor more intrinsically stable coordination modes, e.g., in-phase rather than antiphase patterns for 1:1 frequency-locked modes, and simpler *m:n* frequency lockings such as 1:1 rather than 2:1. Additionally, however, in one formulation of the Haken-Kelso-Bunz (HKB) coupling function, the coupling forces exerted by individual oscillators on one another are subject to a time delay—it can take several cycles before they reach full strength. If we assume that at the fast rate only the 1:1 lowest-order mode is stable enough to resist the effect of fluctuations, then the system will inexorably be drawn to this stable mode as coupling strength increases to its full value. At slower rates, the 2:1 mode may still be relatively stable even at maximal coupling strength, so the buildup in coupling strength over repetitions will not result in such a dramatic increase in

transitions. However, because the 1:1 mode is still the globally strongest attractor state, small fluctuations in the frequency ratio between oscillators may still produce some transitions, even at the slower rates. Thus, the hypothesis that speech production units are dynamic entities that exhibit bifurcations and multistability provides a promising account of the temporal properties observed in this data (buildup and rate-dependence).

Experiment 2: Subsegmental and Segmental Errors—Atomic and Molecular

In our view, segmental units are typically complex in that, while occasionally being isomorphic with a single gestural action unit, they generally entail an assembly of two or more gestural action units into a tightly coupled gestural molecule. For example, English /m/ is composed of a lip closing gesture and a velum lowering gesture that must be produced in a specific coordinative relation. If speech production draws on both a foundational inventory of atomic gestural units and on larger molecular structures, we predict that both individual gestural components of a segment and segmental molecules in their entirety can intrude or reduce in errors under the right coupling conditions.

Methods

For one of seven subjects who participated in Experiment 1, a transducer coil was attached to the velum in addition to the tongue receiver locations described above. Otherwise the same setup as for Experiment 1 was used.

Participants.

Due to the particular difficulties involved in attaching a receiver far back on the palate and the discomfort this causes the subject, data were collected only for one subject and, therefore, only limited data are available. One native speaker of American English with substantial phonetic training participated in this experiment. She was naive as to the purposes of the experiment.

Stimuli and Procedure.

The present utterances were collected in a single session together with Experiment 1, but were presented in a single block before the other consonant conditions. The experimental procedure was the same as described for Experiment 1. Only one rate was employed (112 bpm). Stimuli with alternating nasal and non-nasal final consonants were employed: *bang bad* and *kim kid.* Control utterances were non-alternating (*bang bang, bad bad, kim kim, kid kid*). There was no vowel condition, otherwise the stress and position manipulations of the stimuli were analogous to Experiment 1.[15] Final consonants were chosen rather than initial ones, because the greater extent of velum movement for final compared to initial nasals (Krakow, 1993, 1999) facilitated measurement of velic aperture.

---

[15] The final stress condition trial of *kim kid* had to be excluded from analysis since the subject's initial repetitions were mistakenly of initial stress.

Results

To articulate the dorsal nasal /ŋ/, two oral gestures are hypothesized to be co-active: a raising of the tongue dorsum and, as for all nasals, a lowering of the velum. Likewise, the bilabial nasal /m/ is hypothesized to comprise a lip aperture as well as a velum gesture. For the production of /d/, on the other hand, only a single gesture (tongue tip raising) is hypothesized to be involved. Accordingly, the trajectories measured were vertical positions of the velum, tongue tip, and tongue dorsum for *bang bad*, and velum, tongue tip and lip aperture for *kim kid*. Lip aperture was computed as the Euclidean distance between the upper lip and the lower lip transducer coil.

If gestural units systematically interact in errors as we suggest, we would expect to see two possible error patterns. Both of the gestures participating in producing /ŋ/ might collectively intrude during an intended /d/—that is, an error involving a gestural molecule—or individual gestures might intrude independently during the /d/—that is, an error involving a single atomic action unit. The latter pattern has been referred to as a subsegmental error, though the existence and frequency of this error type has been vigorously debated (see e.g., Dell, 1986; Guest, 2001; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1982). Our data supports the prediction of both subsegmental and segmental gestural errors. While an /ŋ/-like tongue dorsum raising gesture and a velum lowering gesture do sometimes occur together in an erroneous temporal location, there are also cases in which the /ŋ/-like tongue dorsum gesture intrudes *without* an accompanying velum lowering gesture. An example of this kind is shown in Figure 5, which displays

vertical movement of three receivers: (from top) tongue tip, tongue dorsum, and velum. A non-errorful token on the left shows movements for two gestures comprising a velar nasal consonant (/ŋ/): tongue dorsum raising and velum lowering. Neither of these is seen during the tongue tip raising (/d/). In the errorful token on the right, tongue dorsum raising intrudes during tip raising for the /d/ of *bad*, but no velum lowering is seen.

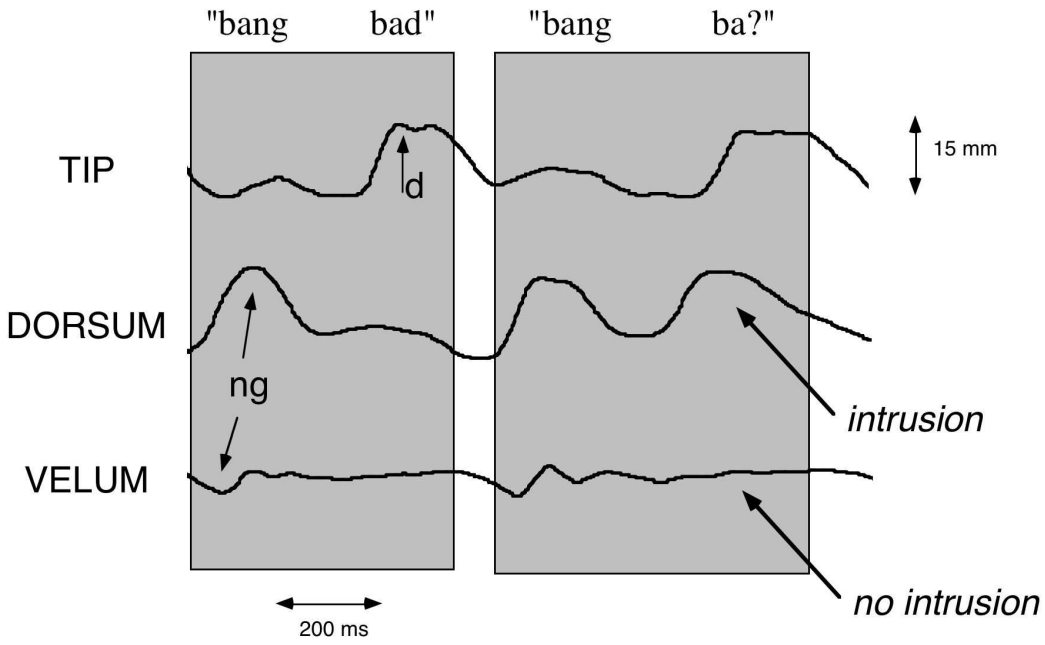

Figure 6. Subsegmental error. Data for two successive repetitions of the phrase *bang bad* (Speaker LK).

The opposite case is also observed: the velum lowering occurs erroneously during the /d/-like tongue tip raising gesture, without the tongue dorsum gesture intruding as well. This results in an /n/-like percept.

The same error evaluation metric as described for Experiment 1 was employed but required several modifications. First, the 75% error criterion that in the earlier experiment led to the exclusion of files from analysis cannot be upheld in a meaningful way; the phrases under consideration here prove to be far more difficult to repeat than the earlier phrases that do not involve nasal consonants. In addition to high error rates, the subject's articulation in the alternating trials differed more from the controls than was the case for Experiment 1 (which this subject also completed). To counteract the elimination of this 75% criterion and to prevent an overestimation of errors, for the following results, only errors that are identified to be of full magnitude by the error metric are counted as error; errors of partial magnitude are disregarded.

When applied to the velum receiver, the error metric proves more problematic. Unlike the points tracked by the tongue and lip transducers, the velum exhibits considerably less movement amplitude in the alternating trials than in the control trials. This bears the consequence that most alternating trials cannot be measured against the controls because all repetitions would be classified as errorful. Looking at the distribution of the alternating trials, some errorful tokens are clearly discernible. Figure 6 gives a scatterplot for the tongue tip and the velum data points obtained during the trial *kim kid,* initial stress. The circles indicate informally the center of the distribution for /m/ (lower left: low tongue tip, low velum) and /d/ (upper right: high tongue tip, high velum).

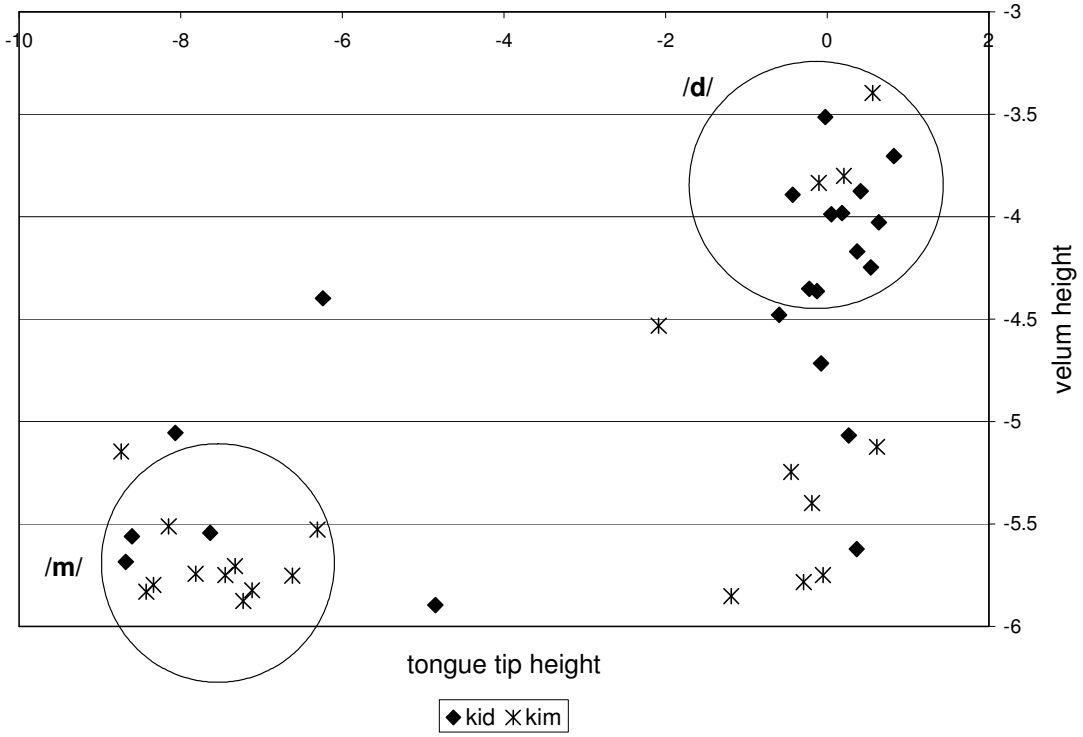

Figure 7. Scatterplot of TT height and velum height measured during the final consonants of the phrase *kim kid*, initial stress. The filled diamonds represent /d/ tokens, the asterisks /m/ tokens.

It can be seen that there are intended /d/s (represented by diamonds) which are clearly in the center of the intended /m/ (represented by asterisks) distribution and vice versa. These tokens cause the distributions of /d/ and /m/ to overlap along the velum height dimension (tongue tip height is plotted for convenience here, it is irrelevant for determining errors in velic aperture). For the velum aperture as indexed by velum height, errors are defined by this overlap criterion; that is, a velum gesture of a magnitude that

makes the distribution of velum heights for /d/ and /m/ overlapping was considered as an error. The minimum number of tokens that can be simultaneously identified in order to achieve separate (non-overlapping) distributions in the velum height dimension are taken to be errorful tokens in terms of velum height. All alternating trials for a given pair of target consonants are considered at the same time. No distinction of partial or full errorful magnitude is made. In some cases, this method led to an ambiguous situation: If, for instance velum height of an /m/ token was just within the velum height variability range of /d/, non-overlapping distributions could also be achieved by marking a /d/ token on the edge of the /d/ distribution as errorful. However, when taking the /d/ token to be the error, this would leave the distributions of the /m/ tokens bimodal (since one /m/ token would be right on the edge of the /d/ distribution while not overlapping it). Here the additional criterion is employed that errorful tokens are to be identified such that the resulting distribution of, for example, the remaining /m/ tokens is not bimodal.

The data are evaluated to answer the question of whether the velum and the respective tongue/lip gestures could individually participate in an error or whether they only move in a holistic fashion. Table 3 gives the overall error numbers obtained for *kim kid* and *bang bad* conditions, respectively.

Table 3

Error Numbers for "kim kid" and "bang bad" Trials.

		gestural errors			
kim kid		velum error	tongue tip error	lip aperture error	number of tokens
intended target	/m/	12	24	5	58
	/d/	15	7	14	58
bang bad		velum error	tongue tip error	tongue dorsum error	number of tokens
intended target	/ng/	9	7	4	56
	/d/	3	3	5	56

All combinations of error co-occurrence are observed: velum errors occur without lingual or labial errors and vice versa.

A chi-square test is employed to test for independence of the velum gesture. The contingency table data for *kim kid* are: Velum error only—11, lip aperture and/or tongue tip error only—21, both—19, neither—65. Contingency table data for *bang bad* are: Velum error only—8, tongue dorsum and/or tongue tip error only—8, both—4, neither—92. The chi-square is significant for both *kim kid* and *bang bad* (*kim kid* chi-squared = 14.908, $p < .01$; *bang bad* chi-squared = 7.1879, $p < .01$), indicating that the probability of a velum gesture error co-occurring with a lip or tongue error is with greater chance , while the probability of a velum gesture participating in an error without its associated

lingual or lip gestures is below chance. Nonetheless, there were trials in which each of the gestures of the gestural molecule individually intruded into another consonant.[16]

Discussion

The results from our experiments provide support for the hypothesis that gestural action units participate in speech production. The observed error patterns conform to the predictions laid out at the end of the Introduction—they involve linguistically significant constriction units, exhibit gradience and patterning that can be understood within a dynamical account of coupled oscillators, and are sensitive to rate (in line with studies showing rate-sensitivity of gestural coupling; cf. Kelso et al., 1986; Krakow, 1999; van Lieshout et al., 1997). Most significantly, many errors were of partial magnitude, and many resulted in an ill-formed structure due to the predominance of intrusion errors.

Despite the systematic occurrence of substitutions and exchanges, a high number of errors were intrusions not accompanied by reductions; this intrusion bias resulted in a grammatically illegal structure. Note that although the outcome of such an intrusion error

---

[16] There is to our knowledge no systematic instrumental investigation of coda consonants in speech errors. While coda errors are reported in transcription studies to be less frequent than onset errors (but see Butterworth & Whittaker, 1980 and Sevald & Dell, 1994), due to the profound differences in methodology to the present study, no meaningful comparisons of error frequency can be made. Whether the present task triggers an unusually high number of coda errors or whether coda errors are de facto underrepresented in transcription corpora has to remain an open question at this point.

was phonotactically ill-formed, these errors were not random distortions. For instance, it was a tongue dorsum gesture typical of /k/ that intruded during the /t/.

The question remains whether the experimental evidence outlined above is of sufficient depth and breadth to rule out an independent role of symbolic or holistic segments in speech production planning (as opposed to 'segments' understood as gestural molecules arising from coupled atomic gestural units). The appeal of a gestural interpretation lies in the possibility that the characteristics of errors as observed in the present experiments, fall out naturally from the characteristics of the hypothesized underlying units of speech production. Partial gestural activations in errors are not expected from abstract segments being serially misordered; however independent mechanisms can be assumed in abstract segment models in order to accommodate these results. Partial gestural activation could for instance be ascribed to a monitoring mechanism or feedback signal that is not fast enough to fully suppress errorful output. Alternatively, it could be assumed that all non-categorical errors happen at a stage in speech production after the symbolic segments have been translated into phonetic representations. While traditional, transcriptional evidence from speech errors is methodologically problematical (as the present study, among other, has shown), recent chronometric data support the role of segments as unanalyzed wholes in speech production planning (Roelofs, 1999). Further research will be necessary to carefully compare the potential roles of atomic dynamical units, gestural molecules, and/or holistic segments in speech production planning.

General Discussion: Tasks and Planning

The most significant, novel findings of our experiments are that during a repetition task, many errors involve the production of an "extra," erroneous speech unit simultaneously with the intended one and that both the intended and intruding units can vary over a range of magnitudes. An intrusion bias in errors has previously been identified in transcription studies (Butterworth & Whittaker, 1980; Stemberger, 1991; Stemberger & Treiman, 1986), yet these studies do not identify errors as *simultaneous* productions of the intended and the intruding segment but as errors creating a *sequential* consonant cluster. We now consider how this novel error type (co-production of errorful and intended units) could arise in other theories of speech production. Most other theories sharply distinguish between speech planning (*phonological encoding*) and the execution of the plans (*articulation*); thus, we can examine models of both of these processes to determine whether one or the other type of model could, as currently constituted, provide accounts of such errors. Theories of phonological encoding (e.g., Dell, 1986; Dell et al., 1993; Levelt et al., 1999; Shattuck-Hufnagel, 1979) differ in the kinds of processes that they hypothesize, but they all share a common output—a linear string of abstract segments assigned to temporal slots in a prosodic frame.

In order for these theories to directly model the co-produced intrusions as planning errors, some deep changes would have to be made to allow two segments sharing a single timing slot to be output. It is not clear what the (potentially undesirable) consequences of such a change would be. Alternatively, maintaining the assumption that the output of the phonological planning stage is a linear sequence of segments, the

simultaneous production of the intended and intruding consonant as well as any gradient movement amplitude observed in errors could be to be assigned to the implementation (execution) processes following the phonological planning stage. Yet why errors would have the form we observe in our experiments is generally outside the scope of phonological planning models. We have offered an account of these errors in terms of dynamical action units (which we take to be units of both planning and execution) 'slipping' into inappropriate coordinations. The account explains why errors of this form occur, and why they should build up over time.

With regard to models of execution (e.g., Perkell, Guenther et al., 2000; Perkell, Matthies, Svirsky, & Jordan, 1995), these typically hypothesize that the speech production process involves navigating a smooth trajectory through articulatory space that satisfies the sequence of segmental goals provided by the phonological plan. It is not clear how such a process would result in an extra "stop" along the path being produced. Indeed, most theorists view the process as one that minimizes articulatory effort (e.g., Lindblom, 1990), a description at odds with intrusive errors (Pouplier, 2003a). Again, modifications would have to be made.

It is conceivable that some combination of changes to a model of planning and/or a model of execution could yield the kind of errors that we have documented while retaining the fundamental supposition that the planning mechanism operates using abstract symbolic segmental units. However, a problem in arriving at candidate changes is that none of the well-developed models (apart from the gestural model we describe) incorporate *both* planning and actual articulation in an integrated or a principled way. For

this reason, it is difficult to evaluate what a change to the planning mechanism would require of the articulation mechanism and vice versa in these other models. For example, planning models could be modified to allow the intended and intruded segments to be output sequentially, if the execution mechanism could be relied on to co-produce them (in most cases). There are simply too many alternatives to consider when planning and execution are treated as independent "modules." Indeed, there has been no principled account within these other type of models of the systematic properties of errors other than those that can straightforwardly be interpreted as serial misorderings of symbolic units. The present gestural account, however, predicts the properties of errors observed in our experiments from the independently motivated architecture of integrated gestural planning and implementation.

The results presented here can be interpreted as providing strong evidence for the deployment of dynamical action units. While the repetitive speech production task employed here is quite different from normal speech production, it seems unlikely that these units would come into being solely as a response to the task demands. Nonetheless, it is possible that the task in some way exaggerates the significance of these units compared to their role in more normal speech production contexts. To test this, a parallel experiment was undertaken using an alternate error induction task that does not involve repetition but rather relies on the visual presentation of primes that tend to provoke errors on an immediately following test trial, that is, the SLIP technique (Motley & Baars, 1976). The results (described elsewhere Pouplier, 2003b, 2005) show that gradient gestural intrusion errors of precisely the kind reported above are also commonly observed

in this task that does not involve any overt repetition. As in the present experiment, the intrusion bias, resulting in ill-formed errors,[17] emerges as a statistically significant effect in this task that lacks any repetition component. Additionally, the repetition task makes minimal demands on planning and might be thought, therefore, to be more revealing of units in execution rather than those involved in speech planning. However, the results from the SLIP experiment (Pouplier, 2003b, 2005) further show that gradient gestural intrusion can appear on the first consonant of an experimental trial, arguably before the dynamics of execution have become established. Further, individuals anecdotally notice that for the task we have presented, errors can "occur in one's head" in silent speech lacking any vocal tract movement; this also lends support to the role of these dynamics in speech planning (cf. also Dell & Repka, 1982; Postma & Noordanus, 1996 on speech errors in inner speech). So while, as with many laboratory speech experiments, we cannot confirm the complete generalizability of our findings to natural non-laboratory speech, we can minimally state with confidence that gradient, gestural (subsegmental) errors are observed in both this repetitive and in the non-repetitive task designed to elicit speech errors (Pouplier, 2003b, 2005). We further speculate, based on these two studies, that

---

[17] It should be pointed out that a gestural intrusion does not necessarily result in a phonotactically illegal gestural configuration. Intrusion of a velum gesture during /d/, for instance, would result in an /n/-like structure. Cf. also Pouplier (2003b) and Pouplier & Goldstein (2005) for the intrusion bias leading to a palatalization bias in /s - sh/ interactions.

these errors and the oscillatory dynamics that underlie them are occurring at the speech planning level (Goldstein, Byrd, & Saltzman, in press; Saltzman et al., to appear), rather than purely at the level of low-level articulatory execution.

Conclusions

In the past, the nature of the compositional units proposed for spoken language has largely diverged from the types of control units pursued in the domains of other skilled motor tasks. A classic source of evidence as to the units structuring speech has been patterns observed in speech errors—"slips of the tongue." We have presented kinematic evidence that dynamical action units—vocal tract constriction gestures—are deployed during the speech production process, while also allowing for the possibility that units of other granularities, which we understand as gestural molecules, are active as well. In the induced speech errors that we have analyzed, gradient gestural intrusions are the most commonly observed error type. Qualitative and quantitative aspects of the error patterns can be explained by general dynamical principles, particularly those governing the behavior of coupled nonlinear oscillators. While the highly structured character of language exerts forces and constraints on speech production units likely to go beyond those on action units for many other motor tasks, this study of speech articulation shows that it is possible and advantageous to develop a principled account of spoken language within a more general theory of action.

ACKNOWLEDGEMENTS

The first two authors have contributed equally to this paper and are in alphabetical order. This paper is dedicated with great fondness and respect to the late Vicki Fromkin who shaped the study of speech errors, and inspired so many of us. We gratefully acknowledge Donca Steriade, Stefanie Shattuck-Hufnagel, anonymous reviewers, and the support of NIH grants HD-01994 and DC-0663 (Haskins Laboratories) and DC-03172 (USC).

REFERENCES

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America, 63*(5), 1535-1555.

Boersma, P. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.

Boucher, V. J. (1994). Alphabet-related biases in psycholinguistic enquiries: considerations for direct theories of speech production and perception. *Journal of Phonetics, 22*(1), 1-18.

Browman, C., & Goldstein, L. (1986). Towards an Articulatory Phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*(2), 201-251.

Browman, C., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica: International Journal of Speech Science, 49*, 155-180.

Browman, C., & Goldstein, L. (1995). Dynamics and Articulatory Phonology. In T. v. Gelder (Ed.), *Mind as Motion. Explorations in the Dynamics of Cognition.* (pp. 175-194). Cambridge, MA: MIT Press.

Browman, C., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée, 5*, 25-34.

Butterworth, B., & Whittaker, S. (1980). Peggy Babcock's relatives. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior* (Vol. 1, pp. 647-656). Amsterdam: North-Holland.

Byrd, D. (1996). A phase window framework for articulatory timing. *Phonology, 13*, 139-169.

Chen, L. (2003). *The Origins in Overlap of Place Assimilation.* Paper presented at the XXIIth West Coast Conference of Formal Linguistics., San Diego, CA.

Cutler, A. (1981). The reliability of speech error data. *Linguistics, 19*, 561-582.

Dell, G. (1984). Representation of Serial Order in Speech: Evidence From the Repeated Phoneme Effect in Speech Errors. *Journal of Experimental Psychology: Learning, Memory and Cognition., 10*(2), 222-233.

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

Dell, G., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149-195.

Dell, G., Reed, K., Adams, D., & Meyer, A. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(6), 1355-1367.

Dell, G., & Repka, R. J. (1982). Errors in inner speech. In B. J. Baars (Ed.), *Experimental Slip and Human Error: Exploring the Achitecture of Volition* (pp. 237-262). New York: Plenum Press.

Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research, 20*(2), 105-122.

Fowler, C., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language Production. Volume1: Speech and Talk* (pp. 373-420). London: Academic Press.

Frisch, S., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30*, 139-162.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.

Fromkin, V. A. (Ed.). (1973). *Speech Errors as Linguistic Evidence*. The Hague: Mouton.

Garnham, A., Shillock, R. C., Brown, G. D. A., Mill, A. I. D., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19*, 805-817.

Goldrick, M., & Blumstein, S. (in press). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*.

Goldstein, L., Byrd, D., & Saltzman, E. (in press). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. Arbib (Ed.), *From action to language: The mirror neuron system*.

Goldstein, L., & Fowler, C. (2003). Articulatory Phonology: A phonology for public language use. In A. Meyer & N. Schiller (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 159-207). Berlin: Mouton de Gruyter.

Guest, D. J. (2001). *Phonetic features in language production: An experimental examination of phonetic feature errors.* PhD dissertation, Urbana-Champaign, Illinois.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*(51), 347-356.

Haken, H., Peper, C. E., Beek, P. J., & Daffertshofer, A. (1996). A model for phase transitions. *Physica D*(90), 176-196.

Kelso, J. A. S., Scholz, J. P., & Schöner, G. (1986). Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. *Physics Letters A, 118*(6), 279-284.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 812-832.

Krakow, R. A. (1993). Nonsegmental Influences on Velum Movement Patterns: Syllables, Sentences, Stress and Speaking Rate. In M. K. Huffman & R. A. Krakow (Eds.), *Nasals, Nasalization, and the Velum*. San Diego: Academic Press.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics, 27*, 23-54.

Kupin, J. (1979). *Tongue twisters as source of information about speech production.* Unpublished doctoral dissertation, University of Connecticut, Storrs, CT.

Laver, J. (1979). Slips of the tongue as neuromuscular evidence for a model of speech production. In H. W. Dechert & M. Raupach (Eds.), *Temporal Variables in Speech. Studies in Honour of Frieda Goldman-Eisler* (pp. 21-26). The Hague: Mouton.

Levelt, W. (1989). *Speaking. From Intention to Articulation.* Cambridge, MA: MIT Press.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1-75.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 217-246). New York: Springer.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of H and H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403-439). Dordrecht: Kluwer.

MacKay, D. (1971). Stress pre-entry in motor systems. *The American Journal of Psychology, 84*(1), 35-51.

Meringer, R., & Mayer, K. (1895). *Versprechen und Verlesen: Eine psychologisch-linguistische Studie*. Stuttgart: Göschensche Verlagsbuchhandlung.

Meyer, A. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. *Cognition, 42*, 181-211.

Motley, M. T., & Baars, B. J. (1976). Laboratory induction of verbal slips: A new method for psycholinguistic research. *Communication Quarterly, 24*(2), 28-34.

Mowrey, R. A., & MacKay, I. R. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*(3), 1299-1312.

Peper, C. E., & Beek, P. J. (1999). Modeling rhythmic interlimb coordination: The roles of movement amplitude and time delays. *Human Movement Science, 8*, 263-280.

Peper, C. E., Beek, P. J., & van Wieringen, P. C. W. (1995). Multifrequency coordination in bimanual tapping: asymmetric coupling and signs of supercriticality. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1117-1138.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer (EMMA) systems for transducing

speech articulatory movements. *Journal of the Acoustical Society of America, 92*, 3078-3096.

Perkell, J., Guenther, F., Lane, H., Matthies, M., Perrier, P., Vick, J., et al. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss.

Perkell, J., Matthies, M., Svirsky, M., & Jordan, M. (1995). Goal-based speech motor control: a theoretical framework and some preliminary data. *Journal of Phonetics, 23*, 23-25.

Postma, A., & Noordanus, C. (1996). Production and Detection of Speech Errors in Silent, Mouthed, Noise-Masked and Normal Auditory Feedback Speech. *Language and Speech, 49*(4), 375-392.

Pouplier, M. (2003a). The dynamics of error. *Proc. XVth ICPhS, Barcelona, Spain*, 2245-2248.

Pouplier, M. (2003b). *Units of phonological encoding: Empirical evidence.* PhD dissertation, Dissertation Abstracts International (AAT 3109449).

Pouplier, M. (2005). Tongue kinematics during utterances elicited with the SLIP technique. Manuscript submitted for publication.

Pouplier, M., & Goldstein, L. (2005). Asymmetries in the perception of speech production errors. *Journal of Phonetics, 33*, 47-75.

Roelofs, A. (1999). Phonological Segments and Features as Planning Units in Speech Production. *Language and Cognitive Processes, 14*(2), 173-200.

Saltzman, E., & Byrd, D. (1999). *Dynamical simulations of a phase window model of relative timing.* Paper presented at the 14th International Congress of the Phonetic Sciences, New York.

Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., & Rubin, P. (1998). Dynamics of intergestural timing: a perturbation study of lip-larynx coordination. *Experimental Brain Research, 123 (4)*, 412-424.

Saltzman, E., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

Saltzman, E., Nam, H., Goldstein, L., & Byrd, D. (to appear). The distinctions between state, parameter and graph dynamics in sensorimotor control and coordination. In A. Feldman (Ed.), *Progress in Motor Control: Motor Control and Learning over the Life Span*. New York: Springer.

Sevald, Ch. & Dell, G. (1994). The sequential cuing effect in speech production. *Cognition, 53*, 91-127.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Lawrence Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 109-136). New York: Springer.

Shattuck-Hufnagel, S., & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41-55.

Stemberger, J. (1982). The nature of segments in the lexicon: Evidence from speech errors. *Lingua, 56*, 235-259.

Stemberger, J. (1991). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language, 30*, 161-185.

Stemberger, J., & Treiman, R. (1986). The internal structure of word-initial consonant clusters. *Journal of Memory and Language, 25*, 163-180.

van Lieshout, P. H. H. M., Hulstijn, W., Alfonso, P., & Peters, H. F. M. (1997). Higher and lower order influences on the stability of the dynamic coupling between articulators. In P. H. H. M. van Lieshout (Ed.), *Speech Production: Motor Control, Brain Research and Fluency Disorders* (pp. 161-170). Amsterdam: Elsevier Science.

Wilshire, C. E. (1998). Serial order in phonological encoding: An exploration of the 'word onset effect' using laboratory-induced errors. *Cognition, 68*, 143-166.