ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Phonetics 33 (2005) 47–75

Journal of Phonetics

www.elsevier.com/locate/phonetics

Asymmetries in the perception of speech production errors

Marianne Pouplier[a,b,*,1], Louis Goldstein[a,b]

[a] *Yale University Linguistics Department, 370 Temple St, P.O. Box 208366, New Haven, CT 06520, USA*
[b] *Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA*

Received 17 April 2002; received in revised form 28 January 2004; accepted 15 April 2004

Abstract

Psycholinguistic research concerned with the mental reality of linguistic units has long relied on speech error data which are traditionally collected by means of impressionistic transcription. Evaluation of these data has been taken to support the view that in word form encoding, the most common form of speech error originates from a categorical mis-selection that shifts a segment to a wrong position within a prosodic 'frame.' Asymmetric distributions in such segmental speech errors have been used to argue for coronal underspecification. However, several relatively recent studies investigating speech errors instrumentally have challenged these assumptions by showing that speech errors are not confined to a categorical position-exchange of segmental units. Specifically it has been shown that the gestures that compose a segment may intrude individually and show up in an incorrect temporal position with variable articulatory magnitude. The overall observed bias for gestural intrusion as opposed to reduction has the consequence that often two gestures (one appropriate, one intruding) are produced simultaneously. The current study tests the perceptual consequences of these phonologically ill-formed errors by presenting listeners with utterances collected in an EMMA speech error experiment. Results indicate that biases in the perception of the ill-formed errors may be the source of asymmetries in error distributions as they have been observed in speech error corpora. Specifically claims about coronal underspecification that have been made on the basis of data collected through transcription are not supported by our study.

[*] Corresponding author. Present address: TAAL, University of Edinburgh, Adam Ferguson Building, 40 George Square, Edinburgh, EH8 9LL, UK. Tel.: +44-131-6503961; fax: +44-131-6503962.

E-mail addresses: pouplier@ling.ed.ac.uk (M. Pouplier), louis.goldstein@yale.edu (L. Goldstein).

[1] Now at the University of Edinburgh, UK. Revisions to this paper were done while the author was at the Vocal Tract Visualization Laboratory, University of Maryland, Baltimore, USA.

0095-4470/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2004.04.001

1. Introduction

Research on speech production has often made use of speech errors in order to gain insight into the functional structure of mental processes. This approach builds on the general notion that constraints that shape the form of minor malfunctions such as speech errors also guide the production of error-free speech. The rationale for this assumption is quite simple: errors are systematic in their occurrence and distribution, as has been noted time and again (cf., for instance, Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979; Dell, 1986; Stemberger, 1991a). This fundamental characteristic of speech errors enables a chain of logical inferences: Those units that appear to behave independently in speech errors are presumably—at least at some point—units of processing. For instance, single segments frequently change places in errors, as for example in *__b__udget __g__ap* turning into *__g__udget __b__ap* (MIT corpus in Shattuck-Hufnagel, 1983), but entire syllables (almost) never directly participate in errors, making the error like *__gui__tune my _tar* from *tune my __gui__tar* (Shattuck-Hufnagel, 1979) extremely unlikely. This has been taken to mean that segments must be direct processing units of word form retrieval, while syllables potentially only play an indirect role in speech production.

At the same time it has just as often been noted that there is a caveat to claims made on the basis of speech error corpora, since some errors are more easily heard than others (due to the split attention situation for the transcriber and general perceptual biases) and can thus come to be overrepresented in corpora (cf. among others Fromkin, 1971; Cole, 1973; Tent & Clark, 1980; Cohen, 1980; Cutler, 1981; Ferber, 1991). While speech error elicitation experiments have confirmed results obtained with spontaneous corpora (Shattuck-Hufnagel, 1983; Stemberger, 1991a), it has to be kept in mind that the method of detecting errors is the same in both cases: impressionistic transcription. The inherently segmental nature of transcription carries its own bias into the data: Subphonemic errors or errors resulting in a phonologically ill-formed utterance, for instance, are difficult to transcribe in a segmental system, which may be one of the reasons why this kind of error is so rarely reported. Mowrey and MacKay (1990), for instance, report that smaller units than phonemes and features regularly participate in errors. Even for audible mispronunciations, the authors find that most of the errors they identified in their EMG data "cannot be characterized alphabetically" (p. 1307).

In recent years, studies have begun to investigate speech errors by means of instrumental measurements. Articulatory and acoustic studies (Mowrey & MacKay, 1990; Boucher, 1994; Frisch & Wright, 2002; Pouplier, 2003; Goldstein, Pouplier, Chen, Saltzman, & Byrd, submitted) have supported the notion of 'gradience' in speech errors empirically, meaning that errors are not a matter of all or nothing—a result which will be explained in detail in the course of this paper. Crucially, the nature of the observed gradience or "partialness" is such that systematic errors can occur below the level of a segment instead of being confined to a temporal misselection of (abstract) phonological units. On the basis of kinematic evidence, Goldstein et al. (submitted) and Pouplier (2003) have demonstrated that speech errors are not restricted to categorical substitutions of segmental units, but rather individual gestures (in the sense of Articulatory Phonology, Browman & Goldstein, 1989, 1992) can exhibit errors that vary from zero to maximal in magnitude.

While the existence of partial speech errors has been established by acoustic measurements (Frisch & Wright, 2002), there has been no systematic examination of the perceptual consequences of errors of varying articulatory magnitude. Goldstein et al.'s (submitted) and Pouplier's (2003) studies enable us to test the perceptual consequences of speech errors whose articulatory properties are known, and thus allow us to assess the magnitude of the perceptual biases that may find their way into transcribed data. We report results from two perceptual experiments which use stimuli selected on the basis of their articulatory properties only, covering a range of errorful gestural activations.

The paper is organized as follows: First, some claims that have been made in the past about the qualitative and quantitative distribution of speech errors shall be summarized together with the theoretical interpretations these data have received. We then discuss how these results have been challenged by recent studies by Goldstein et al. (submitted) and Pouplier (2003) which tracked tongue movement during speech production errors. The kinematic data they obtained will be briefly discussed, since their data served as stimuli in the perceptual experiments we report below. Subsequently, the two perceptual experiments that are the main focus of this paper will be presented. The outcome of these perceptual experiments suggests that different segments show different degrees of vulnerability to (gradient) speech errors: While listeners detected errors reliably for some segments, for other segments the reaction to errorful and non-errorful tokens was not distinct. Observed frequency biases as they have been established on the basis of transcribed corpora can thus be re-evaluated in the light of the perceptual biases found in our experiments: The data suggest that at least for some error types an asymmetric error distribution arises due to perception, while production itself is not asymmetric. However, for errors involving segments whose gestural compositions stand in a subset relationship to each other (as will be explained below), asymmetries may indeed originate in production due to the overall dominance of a gestural intrusion bias observed in the production data.

2. Asymmetric patterns in error distributions

It has been observed throughout the history of speech error research that certain units are more likely to be affected by error than others. The most commonly occurring sound errors are single segment errors, while single feature errors are rarely reported (cf., for instance, Fromkin, 1971; Shattuck-Hufnagel, 1979, 1983; Dell, 1986; Meyer, 1992 for an overview article). An example for a single feature error is given in Shattuck-Hufnagel (1983), where "ti***p*** of the to***ngue***" is reported to have been pronounced as "ti***k*** of the tu***m***." In this error, the respective place of articulation features of the final consonants have switched places, but the voicing and nasality features have not participated in the exchange. Likewise, syllables are usually not directly affected by speech errors.[2] In conjunction with the rarity of sub-segmental errors in speech error corpora, the notion that phonological segments are primary units of word-form retrieval has gained wide-spread acceptance.

[2] Syllable constituents seem to play a role in speech errors in that errors are usually faithful to syllable position, yet whole syllables themselves do not seem to be substituted or exchanged (cf. e.g., Nooteboom, 1973; Dell, 1986; cf. also Shattuck-Hufnagel (1992) for discussion of a position similarity constraint). Likewise, featural similarity plays a role in determining which units are likely to interact, yet single feature exchanges are rarely reported (cf. Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979; Dell, 1986).

The hypothesis that errors must happen at a phonological or cognitive level, as opposed to a phonetic or motor output level primarily rests on the finding that errors are phonologically well formed: An activated (abstract) segment is categorically shifted to a wrong position within a 'prosodic frame.' In this new position, the segment will be produced 'normally', as if it were the intended segment; thus allophonic features typically pertain to the new ('wrong') position of a segment (Shattuck-Hufnagel & Klatt, 1979; Shattuck-Hufnagel, 1983). For instance, while the word initial /p/ in *slumber* ***p***$^{\boldsymbol{h}}$*arty* is aspirated, in the error *_lumber* ***sp****arty* (Fromkin, 1973), the /p/ was heard to be unaspirated, as appropriate for its new position. Also a (hypothetical) error like ***n****arrow* ***g****ap* turning into ***d****arrow* ŋ*ap* is expected to occur only very rarely, since the change of the nasality feature constitutes a phonotactic violation: a velar nasal is not permitted in onset position in English.

Stemberger (1991a,b) has advanced further theoretical claims about units of phonological encoding in terms of underspecification: Errors show a frequency bias in that more frequent elements are less likely to be affected by error than less frequent elements. Relatedly, frequency of occurrence is reflected in a directionality effect: less frequent elements are usually replaced by more frequent elements (cf. also Motley & Baars, 1975). For some segments, however, an anti-frequency bias has been reported (in experimental studies as well as in natural corpora). /t/ is a more frequent segment in English than /k/, for instance, and thus would be expected to replace /k/ more frequently than vice versa, yet the opposite is actually observed: /t/ is more often substituted by /k/. This is also the case for /s/ and /ʃ/, with a /s/ as the more frequent element turning more often into /ʃ/ than vice versa. At the same time, Stemberger and Treiman (1986) and Stemberger (1991a) identify an addition bias in cluster environments: In errors, it is more usual for a segment to be added than to be deleted. Stemberger and Treiman found in an error elicitation experiment that for instance, on the word pair *puck plump* the most frequent error was *pluck plump*, while *puck pump* occurred only rarely. They summarize the dominant pattern as "something" competing with "nothing", with the consequence that the "something" wins (Stemberger & Treiman, 1986; cf. also Stemberger & Stoel-Gammon, 1991).

Stemberger (1991a) reinterprets the anti-frequency effect as surface manifestation of the addition bias by invoking the concept of coronal underspecification. Given that coronals are underspecified for place of articulation, so he argues, the addition bias will lead other segments' place specifications to intrude more easily (since the 'empty space' is a willing host), independent of segment frequency. Again, "something" competes with "nothing," that is, the anti-frequency effect is conceptualized as a consequence of an addition bias at the featural level.

Stemberger's results are contradicted by a recent study of Goldstein et al. (submitted) and Pouplier (2003) which tracked tongue movement during error elicitation experiments. On the basis of these kinematic data the authors found that in an error, components of gestural structures can be activated to varying degrees. That is, both individual gestures as well as larger units consisting of tightly cohesive multiple gestures can be involved in erroneous productions. The authors observe a systematic bias for gestural intrusion as opposed to reduction with the consequence that often two gestures are produced at the same time (one appropriate, one intruding). For example, during the repetition of the phrase *cop top*, errors are observed in which an intruding /k/-like dorsum gesture is produced concurrently with the tongue tip gesture of *top*. However, no asymmetry is found between /t/-like tongue tip gestures and /k/-like tongue dorsum gestures—they are equally likely to intrude during the other consonant.

A potential source for the divergent findings may lie in the different methodology employed. While Goldstein et al. (submitted) as well as Pouplier (2003) identified errors exclusively on the basis of articulation, Stemberger relied on auditory evaluation and transcription. Since phonologically ill-formed errors as found in the articulatory data are not necessarily represented and hardly representable in a notation system that is inherently segmental, the strict reliance on transcription for speech error research might not be appropriate (cf. also Mowrey & MacKay, 1990; Boucher, 1994; Frisch & Wright, 2002). The motivation for the perceptual study reported in this paper is thus twofold: First, to investigate the perceptual consequences of gradient speech errors, and second, to shed light on the relationship between (asymmetric) error distributions and perceptual biases. The current work shows that asymmetries in speech errors can be accounted for without appealing to the concept of coronal underspecification.

The next section will briefly summarize some relevant aspects of Goldstein et al.'s (submitted) and Pouplier (2003) studies since a subset of the kinematic data from two of their experiments served as stimuli in the current perceptual study.

3. Articulatory evidence: kinematic data

Goldstein et al. (submitted) and Pouplier (2003) employed a magnetometer (electromagnetic midsagittal articulometer, EMMA; cf. Perkell et al., 1992) system in a speech error elicitation experiment. Their results for a single female talker were used as stimuli for the perceptual experiments reported below.

In this kinematic study, the subject was instructed to repeat bisyllabic stimuli with alternating onset consonants for several seconds, such as *cop top* or *sop shop*.[3] Control utterances are nonalternating phrases, such as *cop cop* or *shop shop*. During the nonerrorful control utterances systematic tongue tip (henceforth TT) movement towards a constriction goal during /t/ is observed, while during /k/, the tongue dorsum (henceforth TD) movement towards a constriction goal is observed. Fig. 1 exemplifies errors as they were obtained in the alternating conditions. Each picture shows two time functions: one non-errorful, prototypical repetition and one errorful repetition, marked by an arrow. During the TT raising gesture associated with /t/, for instance, no substantial TD raising is expected on the basis of the controls. The raising of TD during /t/, as seen in Fig. 1a is thus defined as intrusion error. During other tokens, anomalous TT raising during /k/ can be observed (cf. Fig. 1b).

Goldstein et al. (submitted) and Pouplier (2003) define *intrusion errors* as addition of a tract variable constriction that is not controlled for in the normal, nonerrorful production. Another type of error is *reduction errors* in the target constriction, marked in Figs. 1c for /t/ and 1d for /k/.

By means of an error metric that is based on distributional characteristics of the nonalternating control utterances, Goldstein et al. and Pouplier determine that for both segments, there are significantly more intrusion than reduction errors. This finding has two important theoretical

[3] For the EMMA experiment the following experimental variables were used: stress (iamb, trochee), rate (fast, medium, slow), phrase position (*cop top* vs. *top cop*) and vowel (*top cop* vs. *tip kip*). For controls, nonalternating phrases were used (*cop cop*, *top top*). The acoustics were recorded with a Sennheiser microphone which was positioned in front of the subject. Sampling rate was 22 kHz. The reader is referred to Goldstein et al. (submitted) and Pouplier (2003) for more details.

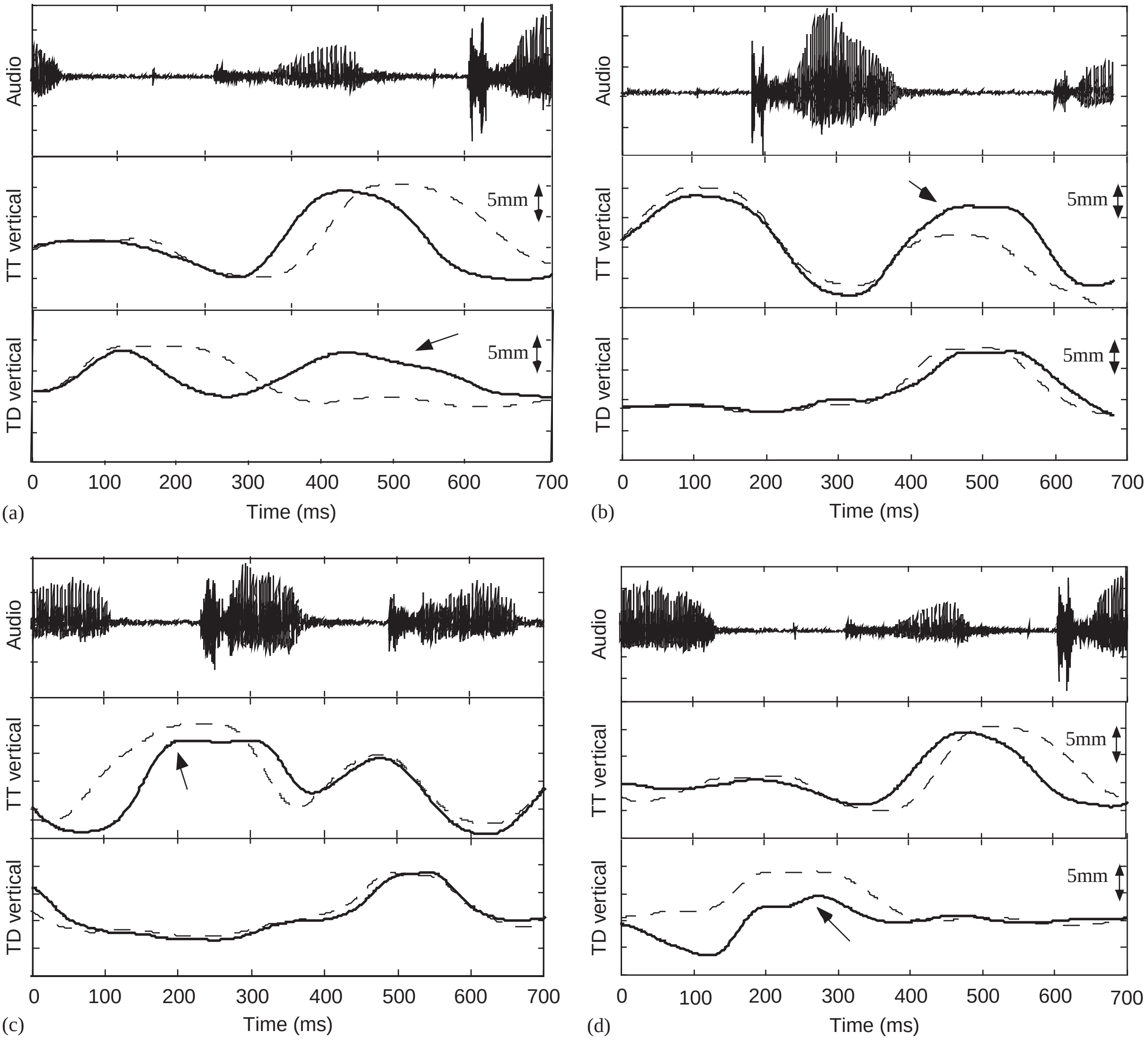

Fig. 1. Examples for intrusion and reduction errors during /t/ and /k/. The dashed line presents a prototypical, error-free repetition. The arrow indicates the error. (a) *cop top*, slow rate, iamb. TD intrusion during /t/. (b) *top cop*, slow rate, trochee. TT intrusion during /k/. (c) *top cop*, mid rate, trochee. TT reduction during /t/. (d) *cop top*, slow rate, iamb. TD reduction during /k/.

implications: First of all, the addition bias, hypothesized by Stemberger (1991a), is confirmed on a gestural level: the addition of an erroneous gesture occurs more often than the reduction of a target gesture. Further, the asymmetry in intrusion vs. reduction shows that errors are not necessarily phonologically well-formed. Since intrusion is frequently not associated with reduction, errors often exhibit a simultaneous co-production of two gestures. In the context of the present paper, it is particularly important to note that the intrusion bias is not confined to gradient errors; it holds across the continuum of magnitudes. Under the assumption that coronals are underspecified for place at the level of the speech production system where substitution errors happen, it would be expected that any existing asymmetry between /t/ and /k/ should only be found for errors whose magnitude is large enough to be analyzed as categorical at the level of gestures. The data collected by Goldstein et al. and analyzed in detail in Pouplier (2003) show no significant asymmetry in error distribution between /t/ and /k/. Instead, a highly significant asymmetry marked by the intrusion bias holds equally for /t/ and /k/ across all error magnitudes.

For the second consonant pair under consideration, /s - ʃ/, errors of gradient magnitude were obtained as well.[4] As to their gestural composition, /ʃ/ is hypothesized to be produced with a tongue tip and a tongue body (henceforth TB) gesture, while /s/ is presumably produced with a TT gesture only (cf. Pouplier, 2003 for a more detailed discussion). For TB, Pouplier (2003) reports an intrusion bias—TB during /s/ intrudes more often than it is reduced during /ʃ/. While both /s/ and /ʃ/ exhibit errors on TT, no meaningful distinction intrusion-reduction can be made, since TT is hypothesized to be actively controlled for in both segments.[5] There is no statistically significant difference in number of TT errors during /s/ as opposed to during /ʃ/.

In sum, the results of the kinematic experiments of Goldstein et al. and Pouplier (2003) show that there is no asymmetry in production between /t/ and /k/—the relevant asymmetry is between reduction and intrusion errors instead. This asymmetry affects /t/ and /k/ equally. For /s/ and /ʃ/, the situation is more complex, since the distinction intrusion–reduction can only be defined in terms of the nonshared vocal tract variable (TB). For TB, the addition bias is confirmed, but in this case results in a /s - ʃ/ asymmetry.

These findings lead to the question where the asymmetry that has been reported between coronals and noncoronals might stem from. Asymmetries may have two potential sources: In some situations (to be discussed in more detail below), the intrusion bias may lead to one segment prevailing over another, that is, asymmetric error distributions may originate in production. As we have just seen, however, distributional asymmetries between /t/ and /k/ are not consistently evidenced in the production data described above. Asymmetries may also originate in perception: The intrusion bias might have different perceptual consequences for different segments. In recording errors, perceptually more salient errors may thus come to be overrepresented. That is, gradient errors and their interaction with perceptual biases might account for the asymmetries. In order to put this possibility to test, the data described above were used as stimuli in a perceptual experiment with the aim of determining how the articulatory error distribution maps onto a distribution of perceived errors.

[4] Due to technical difficulties in the analysis of the kinematic data, only one vowel condition (*sop shop*) is taken into consideration for the /s - ʃ/ pair. These data were obtained in the same experiment as the /t - k/ data of Goldstein et al. (submitted) and are described in Pouplier (2003).

[5] TT can be distinguished for /s/ and /ʃ/ by the fact that TT consistently raises higher for /ʃ/ than for /s/, presumably due to the different shape of the palate at the more posterior constriction location of /ʃ/ (the more posterior constriction location for TT is hypothesized to be due to the high TB).

4. Perceptual experiments

4.1. Experiment 1: /t – k/

4.1.1. Method

The bisyllabic alternating phrases of the articulatory study were edited into single word utterances (e.g., *cop*). In a *go–no go* reaction time task, participants were instructed to listen to "short words" presented in random order and decide whether the words begin with a particular consonant sound. Subjects ranged from 18 to 44 years of age and were paid for their participation. Fourteen subjects were tested; they were all naive as to the purpose of the experiment. All subjects reported normal hearing. Data from 3 subjects were discarded since their identification rate for the error-free controls was below 50%.

The stimulus list contained 71 single syllable tokens, taken from all rate, stress, vowel, and phrase position conditions (cf. Appendix A). Tokens were selected according to two criteria: First, tokens were chosen according to Error Magnitude Category. The error categories are schematically represented in Table 1.

None of the tokens selected has an error on more than one constriction (i.e., never a reduction as well as intrusion error on the same token).[6] Where possible, a minimum of 10 tokens from each category was selected for both /t/ and /k/. The two stimulus categories of exceptions were categorical and gradient reduction errors: The EMMA data for this subject experiment did not contain any reduction errors of categorical magnitude.[7] Also gradient reduction errors, i.e., errors on the target gesture (TD for /k/ and TT for /t/) with no errorful intruding gesture (TD for /t/ and TT for /k/) are underrepresented with only one occurrence for *kip* and *tip*, but 4 during *top* and five during *cop*.[8]

Secondly, within each error group represented in Table 1 (for both /t, k/), tokens with different error magnitudes were selected. For instance, within the category 'gradient error on TD' for *top*, some tokens were selected for which the value of TD was close to what Goldstein et al. (submitted) and Pouplier (2003) define as nonerrorful distribution, while others were close to the nonerrorful distribution of /k/ (i.e., close to being a categorical, not gradient error). In this way, errors covered a range of gestural magnitudes, which then could be correlated with reaction times.

Using the software PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993), the stimuli were presented 12 times overall (six times each in two different conditions), randomized differently each time. Subjects sat in a sound attenuated booth in front of a computer screen and a button box. Stimuli were presented over headphones. Two different monitoring conditions were employed; in one condition subjects were asked to decide whether they heard an initial /t/-sound, in another condition subjects should monitor for an initial /k/-sound. If they heard the given sound, they

[6] Since categorical substitutions are rare in the kinematic data, we were not in a position to systematically include substitution errors in the experiment and still confine the stimuli to a single speaker.

[7] Pouplier (2003) reports only three reduction errors of categorical magnitude across 7 subjects (0.1% of tokens), none of these errors occurred for the subject whose data were chosen for the perceptual study. The data from this particular subject were used because the EMMA data collection for this speaker successfully included all of the experimental variables (phrase position, rate, stress, vowel).

[8] The difference between /t/ and /k/ in the gradient-categorical intrusion categories is due to a coding mistake in the experimental setup; these numbers do not reflect genuine differences in frequency of error.

Table 1
Single syllable stimuli for perceptual task grouped into error categories. Numbers in parentheses indicate the number of tokens representing each category

Error type	Error magnitude category		
	Categorical	Gradient	No error
Intrusion	*t* (10)	*t* (10)	
	k (12)	*k* (8)	*t* (10)
Reduction	—	*t* (5)	*k* (10)
	—	*k* (6)	

were instructed to press a response button as quickly as possible, otherwise, they were instructed to wait for the next trial. The conditions were blocked in cycles of three, that is, the program cycled three times through the entire /t - k/ stimulus list in three different randomizations while subjects monitored for /t/. In the subsequent three cycles, subjects were asked to monitor for /k/, then again for /t/ and once more for /k/. Between these blocks of three cycles, subjects were given the option to take breaks, which they ended by pressing the response button. During a given block, a letter representing the sound subjects should be monitoring for was displayed on the screen. The time between the onset of two successive stimuli was 2000 ms, partitioned into a 1500 ms response window and 500 ms inter-trial time. During the window of 1500 ms subjects heard the audio stimulus and response time was measured (the response window started with the onset of the audio stimulus). Independent of whether a response was recorded, the next trial came up after 2000 ms.

The instructions given to the subjects were as follows:

> You will be hearing short words presented in random order. Your task is to listen and decide whether the words begin with a particular consonant sound. At the start, the instructions on the screen will ask you to listen for a /t/-sound as in ‘top’ or ‘tip.’ If you hear the sound, please press the response button as fast as you can. If you do not hear the sound /t/ in a given trial, do not press the button and just wait for the next trial. You will always see a letter representing the sound you should listen for on the screen.
>
> You will have the possibility to take short breaks during trials. A message will appear on the screen and give you the option to take a rest. When you are ready to continue, you hit the response button.
>
> After a certain number of breaks, instructions on the screen will tell you to listen for a /k/-sound as in ‘cop’ or ‘kip.’ Again, if you hear the sound /k/ in a given trial, press the button as fast as possible, if you do not hear the sound /k/, wait for the next trial. Following that you will be once again asked to listen for /t/ and then for /k/.
>
> At the very beginning, there will be a practice period of 16 tokens separate from the actual experiment.

Table 2
Duration means for /t - k/ stimulus tokens grouped by error magnitude category and rate

Error magnitude category	Rate	Mean	s.d.	*N*
Categorical	Fast	115.1	28.1	13
	Mid	165.4	67	7
	Slow	175.8	n.a.	1
	Total	134.8	49.8	21
Gradient	Fast	116.6	23.8	11
	Mid	110	34.5	12
	Slow	147.1	23	7
	Total	121.1	31.3	30
No error	Fast	107.4	23.6	6
	Mid	122.7	20.1	9
	Slow	180.3	29.9	5
	Total	132.5	36.8	20
Total	Fast	114.1	25.1	30
	Mid	127.9	46	28
	Slow	162.1	29.1	13
	Total	128.3	39	71

The duration of the tokens used as stimuli differs systematically with rate. This opens the possibility that subjects respond systematically to rate differences in the stimuli, as opposed to differences in gestural magnitude. For /t - k/ tokens, the average duration values are given in Table 2.

A two-way ANOVA with the factors Rate (fast vs. mid vs. slow) and Error Magnitude Category (categorical vs. gradient vs. no error) shows that the mean durations differ significantly from each other with rate, with tokens being longest at the slow rate and shortest at the fast rate ($F(2, 62) = 7.845$; $p = 0.001$). The main effect Error Magnitude Category is not significant ($F(2, 62) = 2.268$; $p = 0.112$), nor is the interaction ($F(4, 62) = 2.383$; $p = 0.061$). The interaction shows a small effect because tokens of categorical errorful magnitude are longer at the mid rate than the other error magnitude categories, while at the fast rate, all error magnitude categories are more similar in duration. At the slow rate, tokens with gradient errors are on average shorter in duration than categorical or no errors. Overall, durational cues thus cannot account for any differences between error magnitude categories. Table 3 displays the mean duration by error category for both /t/ and /k/. A further two-way ANOVA tests whether subjects' responses could be influenced by potential durational differences between the coronal and dorsal consonants. The factors were Intended Target (/t/ vs. /k/) and Error Magnitude Category (categorical vs. gradient vs. no error). An intended target is defined as the target the speaker was instructed to pronounce in the production experiment.

None of the main effects nor the interaction are significant (Intended Target—$F(1, 65) < 1$; Error Magnitude Category—$F(2, 65) < 1$; interaction—$F(2, 65) < 1$). Durational cues were thus not sufficient to systematically affect listeners' monitoring behavior.

Table 3
Duration means for /t - k/ stimulus tokens grouped by error magnitude category and intended target

Error magnitude category	Intended target	Mean	s.d.	N
Categorical	/k/	133.5	44.7	11
	/t/	136.2	57.3	10
	Total	134.8	49.8	21
Gradient	/k/	121.6	31.7	15
	/t/	120.6	32	15
	Total	121.1	31.3	30
No error	/k/	127.7	33.4	10
	/t/	137.3	41.1	10
	Total	132.5	36.8	20
Total	/k/	126.9	35.9	36
	/t/	129.8	42.5	35
	Total	128.3	39	71

4.1.2. Results and discussion

The goal of our analysis is to determine how detectability of /t/ and /k/ is affected by error status. Instead of analyzing percent of correct identifications directly, the data are transformed using d' as sensitivity measure (Macmillan & Creelman, 1991). This sensitivity measure takes into account subjects' inherent response bias by adjusting the number of hits (i.e., correct identification responses) for the number of false alarms (i.e., incorrect positive responses). d' is computed by converting hit and false alarm proportions to z-scores and subtracting false alarms from hits; the formula is given in

$$d' = z(H) - z(F), \tag{1}$$

where H is the proportion of hits relative to the number of trials during which a signal is present and F is the proportion of false alarms relative to the number of trials during which no signal is present. Since perfect accuracy (only hits, no false alarms) has an infinite d prime value, proportions of 1 and 0 need to be adjusted to a smaller value. Where $H = 0.99$ and $F = 0.01$, $d' = 4.65$, which is considered as effective ceiling (Macmillan & Creelman, 1991). Proportions of 1 and 0 are thus adjusted to 0.99 and 0.01, respectively.

Goldstein et al. (submitted) and Pouplier (2003) find that the production data are generally characterized by the dominance of intrusion errors. Conceivably, the perception data would follow the same pattern, that is, intrusion of an erroneous gesture might have a greater effect on perception than a reduced magnitude of the target gesture (reduction error). It might also be possible though for the magnitude of the errorful gesture to be the dominant perceptual factor. Due to the fact that there are no full reduction errors in the stimuli, it is not possible to conduct an ANOVA that would allow us to test for interactions between intrusion, reduction and error magnitude category. Instead, a repeated measures ANOVA on the d' data is conducted by partitioning the factor Error Type into four levels: categorical intrusion, gradient intrusion,

Table 4
Means with standard deviations in parentheses for d' results for /t/ and /k/ grouped by error type

	Intended target	
Error type	/t/	/k/
Categorical intrusion	1.51 (0.3)	2.14 (0.19)
Gradient intrusion	2.9 (0.23)	2.22 (0.22)
Gradient reduction	2.77 (0.24)	2.71 (0.27)
No error	4.1 (0.18)	2.43 (0.25)

gradient reduction, no error.[9] The present ANOVA thus has two factors: Error Type with four levels, and Intended Target which has two levels, /t/ and /k/. Intended target is defined as the target the speaker was instructed to pronounce in the production experiment. For each subject, the d' values are calculated across tokens per error type for both /t/ and /k/ (e.g., gradient reduction during /t/, non-errorful /k/, etc.). Each subject thus contributes one d' value per error type for both /t/ and /k/. Table 4 gives the means and standard deviations for the intended targets grouped by error type:

The factor Intended Target does not reach significance ($F(1, 10) = 2.186$, $p = 0.17$), the factor Error Type is significant with $F(3, 30) = 27.332$ ($p < 0.0001$), also the interaction effect is significant ($F(3, 30) = 17.028$, $p < 0.0001$). A one-way ANOVA follow-up conducted separately for /t/ and /k/ with a posthoc test (Ryan–Einot–Gabriel–Welsch Multiple Range Test, $\alpha = 0.01$)[10], designed to test for significant difference in the interaction means, shows that for /k/ none of the error types differ significantly from each other. For /t/, however, no error, gradient error (intrusion and reduction) and categorical error are significantly different from each other. That gradient intrusion and reduction are not significantly different from each other, but both differ significantly from categorical intrusion suggests that for /t/, the primary factor influencing correct identification seems to be error magnitude rather than intrusion vs. reduction.

Although the main effect Target does not reach significance, it should be pointed out that /t/ scores higher d' values for the nonerrorful tokens than /k/. While it is not entirely clear why the average identification index for /k/ does not go above a d' value of 3, and yet scores for the identification of nonerrorful /t/ are close to the d' ceiling value, the crucial point in the current context is the interaction effect. d' values for /k/ are uniform across categories, while the identification of /t/ is clearly affected by the ambiguity introduced through the errorful gesture.

Besides scoring the data for correct answers, we collected response times. For the following analyses, only datapoints are included when subjects had correctly identified an intended target, that is, only hit responses were included. This limits the data somewhat, since for the targets with errors, naturally a higher miss rate is expected than for error-free tokens; thus fewer data points

[9] While tokens were selected from all phrase position, stress and vowel conditions of the articulatory data, these conditions were not tested separately in the analysis of the perception data, since priorities in token selection were on the factors gestural magnitude, error magnitude category, and error type.

[10] REGWQ is a modified Newman–Keuls test, also based on the q-statistic (cf. Toothacker, 1993, 38ff.). Toothacker recommends the use of REGWQ by virtue of its good alpha control compared to, for instance, Newman–Keuls while being a more powerful statistic than, for instance, Tukey HSD.

Table 5
Means with standard deviations in parentheses for reaction time results for /t/ and /k/ grouped by error type

	Intended target	
Error type	/t/	/k/
Categorical intrusion	631.76 (84.91)	589.43 (59.57)
Gradient intrusion	580.88 (83.59)	592.23 (56.12)
Gradient reduction	606.47 (81.12)	610.76 (51.02)
No error	563.11 (74.89)	593.52 (59.39)

were recorded compared to error free tokens. There is no token, however, that was consistently mis-identified across subjects, that is, no token has only misses and no hits for all subjects. For a repeated measures ANOVA with reaction times as observation points, reaction times are averaged so that each subject contributes one data point for each error type for each /t/ and /k/ (e.g., gradient reduction during /t/, nonerrorful /k/, etc.). Averaged this way, hit responses (and thus tokens for which the recorded RT entered into the present analysis) range from 62% to 100%, with an average of 79% for tokens with categorical errors, 87% for tokens with gradient errors and 94% for error-free tokens.

We predict for the results that to the extent that errorful tokens are ambiguous, they should lead to slower reaction times in this categorization task. Factors are defined as above for the d' analysis (Intended Target, two levels, and Error Type, four levels). Table 5 displays the cell means.

Results show a similar overall pattern as before: The factor Intended Target fails to reject the null hypothesis ($F(1, 10)<1$). The factor Error Type is significant ($F(3, 30) = 13.838$; $p<0.0001$) as is the interaction ($F(3, 30) = 12.645$; $p<0.0001$). Looking at the means in Table 5 we see that the interaction effect arises for slightly different reasons compared to the d' results. Parallel to the d' results neither the error status of TD nor of TT substantially affects reaction times for /k/. Using a pairwise comparison of the means with the Bonferroni adjustment for multiple comparisons we can determine that for /t/, categorical intrusion and gradient reduction have a significant effect on reaction times, whereas gradient intrusion does not significantly differ from error-free tokens. Compared to the d' results this means that while gradient intrusion and reduction are similar with respect to correct identification ratios, subjects are slower in identifying the intended target when the target gesture is reduced compared to cases where a gesture partially intrudes. Overall, the reaction time data confirm that articulatory ambiguity affects the perception of /t/, but has no systematic effect on the perception of /k/.

A correlation analysis is carried out in order to establish whether there is a predictable relationship between absolute gestural magnitude (in mm) and reaction time. Again, only hit responses were analyzed.[11] The following patterns are predicted: For TD during /k/, the target gesture for the production of the velar stop, a negative correlation was expected: The higher the TD, the faster the response; likewise, a reduced target gesture will slow down reaction times. For

[11] Since the experiment is a *go–no go* design there are not enough data points to make statistically grounded observations about the reaction times for false alarms, that is, cases in which for instance subjects identified an intended, albeit errorful /t/ as /k/. For the datapoints that are available, reaction times for false alarms are consistently slower across error categories than for hit responses.

Table 6
Correlation results for absolute gestural magnitude and reaction time

Tract variable	Intended target	N	r	p
TD	/t/	370	0.275	<0.0001
TD	/k/	377	0.021	0.686
TT	/t/	370	−0.26	0.623
TT	/k/	377	0.03	0.558

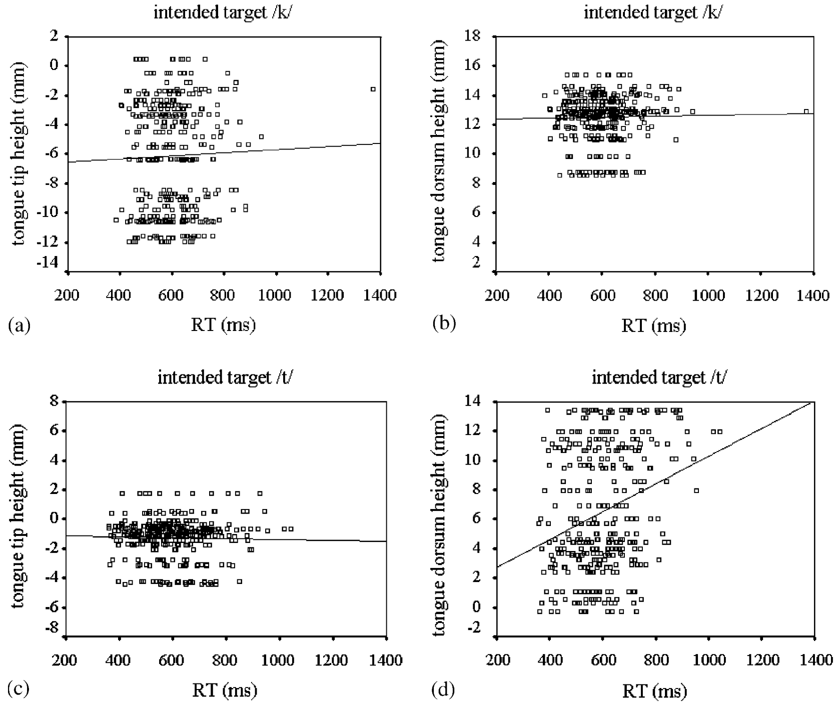

Fig. 2. Scatter plots for correlations between gestural magnitude and reaction times (hit responses only). Each data point represents the averaged responses for multiple presentations of the same token within a given subject. (a) Correlation TT height—response time for intended target /k/. (b) Correlation TD height—response time for intended target /k/. (c) Correlation TT height—response time for intended target /t/. (d) Correlation TD height - response time for intended target /t/.

TD during the production of a /t/, the opposite should be the case: the higher the TD, the higher reaction times should be; we predict a positive correlation. For TT, we again anticipate a negative correlation where TT is the target gesture, but a positive correlation where TT is an intruding gesture. Responses are averaged for multiple presentations of the same token within a given subject.[12] Table 6 shows the results of the correlation analysis. Fig. 2 shows the corresponding scatter plots.

[12] Averaged this way, 34 tokens out of 781 (11 subjects × 71 tokens) have a zero hit rate. This means there are 34 instances in which an individual subject did not respond to any of the multiple representations of a given token.

The results have to be interpreted with care due to the small correlation coefficients, but they mostly pattern with the above results of the sensitivity analysis: The effect for /t/ is in the predicted direction, the target gesture (TT) shows a positive correlation and the intruding TD gesture shows a negative correlation, but only the latter reaches significance. Response times for /k/ do not correlate significantly with gestural magnitude, and the correlation is not in the predicted direction when correlating TD height with response time (note though that *r* is extremely small). Although the outcome of this analysis is comparatively weak, it provides supporting evidence for the perceptual asymmetry between /t/ and /k/ that became evident in the *d′* analysis.

As to the apparent anti-frequency bias that has been reported for /t/ and /k/ for error data collected by means of transcription, the present results indicate that these asymmetries may reflect a property of the perceptual system rather than the production system: if errors are systematically heard more easily on /t/ than on /k/, this perceptual asymmetry can be expected to substantially affect the error distribution in corpora.

4.2. Experiment 2: /s - ʃ/

4.2.1. Subjects and experimental setup

The same experimental setup and the same subjects (in a separately scheduled session) were used for /s - ʃ/; 14 of the earlier subjects were available. For Experiment 2, data from 2 subjects were discarded since their identification rate for the error-free controls was below 50%.

For /t - k/, stimuli had been selected such that there were never two co-occurring errors during one token, that is, no token had an error on both constrictions at the same time. For /s - ʃ/ this selection criterion has to be modified since TT and TB are presumably not independent in the way TT and TD are (TT and TB receivers were about 20 mm apart). That is, it is not possible to select sufficient tokens with an error on TT only or TB only; for most tokens, an error on TT is accompanied by an error on TB and vice versa (cf. Pouplier, 2003). Nor is it possible to vary error magnitude category (i.e., gradient TT with categorical TB error, gradient TT with categorical TB error, etc.) systematically and cover a range of errorful gestural activations within each category, since not enough representatives of each type are in the available EMMA data. It is thus deemed best to focus on TB in token selection, since by virtue of the intrusion bias TB is the locus of asymmetry in the production data. Note that nonerrorful tokens are truly ‘error free;’ gestural magnitudes for both constrictions are well within the bounds of the nonerrorful distribution.

The stimuli distribution is given in Table 7. Selected tokens are distributed across all rate, stress and phrase position conditions (cf. Appendix B).

Since only the activity of one tract variable (TB) can be differentiated in terms of reduction and intrusion, fewer tokens are tested than in Experiment 1. In addition, only error data for one vowel condition are available from the EMMA experiment. The stimulus list for the fricatives thus contained 29 tokens.

Due to coarticulation, the release of the preceding /p/-closure in a *sop shop* phrase is audible during the frication, even after the utterances have been cut up into individual words. To ensure that subjects would parse the bilabial release as coda instead of as complex /ps/ or /pʃ/ onset, a syllable /op/ was spliced at the beginning of all tokens. The presence of a vowel ensures a coda parsing for /p/, since /ps/ or /pʃ/ are not licit onsets in English. The syllable was a stressed, fast rate, nonerrorful utterance of *shop*: The frication part as well as the first temporal half of the

Table 7
Stimulus categories for /s - ʃ/ according to TB error status. Numbers in brackets indicate the number of tokens representing each category

TB	Error magnitude category		No error
Error type	Categorical	Gradient	
Intrusion	s (5)	s (4)	s (5)
Reduction	ʃ (5)	ʃ (5)	ʃ (5)

Table 8
Duration statistics for /s - ʃ/ stimulus tokens by error magnitude category and rate

Error magnitude category	Rate	Mean	s.d.	*N*
Categorical	Fast	356.4	49.7	3
	Mid	359.8	19.5	3
	Slow	436.1	66.2	4
	Total	389.3	61	10
Gradient	Fast	367.6	31.8	5
	Mid	326.3	6.1	2
	Slow	435.5	20.4	2
	Total	373.5	45.9	9
No error	Fast	346.2	19.5	4
	Mid	365.1	53.9	2
	Slow	430	33.7	4
	Total	383.5	49.8	10
Total	Fast	357.7	31.9	12
	Mid	351.7	30.4	7
	Slow	433.5	43.5	10
	Total	382.4	51.3	29

vowel were cut off (resulting in a vowel duration of 55.1 ms). A silence interval of 100 ms was spliced to the end of the /op/ vowel and formant transitions to simulate the /p/ closure. The instructions specified that subjects would hear bisyllabic words with the first syllable always being /op/. It was specifically pointed out to them that there was no consonant at the beginning of the word in order to avoid confusion with the task of the first experiment. Their task was specified as determining whether the second syllable begins with a given consonant sound.[13]

[13] The following paragraph was added to the instructions given to subjects: "Like last time, you will be hearing short words presented in random order. This time, however, the words have two syllables. For all trials, the first syllable is 'op.' Your task is to listen and decide whether the words have a particular consonant sound at the beginning of the second syllable. There is no consonant at the beginning of the word." Otherwise, the instructions were, mutatis mutandis, the same as for Experiment 1.

Table 9
Duration statistics for /s - ʃ/ stimulus tokens by error magnitude category and intended target

Error magnitude category	Intended target	Mean	s.d.	*N*
Categorical	/s/	377.9	39.4	6
	/ʃ/	406.4	89	4
	Total	389.3	61	10
Gradient	/s/	396.8	68.5	3
	/ʃ/	361.9	31.6	6
	Total	373.5	45.9	9
No error	/s/	362.5	36.9	5
	/ʃ/	404.6	55.8	5
	Total	383.5	49.8	10
Total	/s/	376.4	43.7	14
	/ʃ/	388	58.6	15
	Total	382.4	51.3	29

For /s - ʃ/, the durational distribution of tokens grouped by error magnitude category and rate is given in Table 8. Table 9 displays mean token duration grouped by error magnitude category and intended target.

A two-way ANOVA with the factors Rate (fast vs. mid vs. slow) and Error Magnitude Category (categorical vs. gradient vs. no error) shows that the stimulus tokens differ significantly with rate ($F(2, 20) = 12.406$, $p < 0.0001$), with token duration increasing with decreasing rate, but neither the main effect Error Magnitude Category ($F(2, 20) < 1$), nor the interaction ($F(2, 20) < 1$) reach significance. As for /t - k/, a further two-way ANOVA tests whether the sibilants differ systematically in duration and could thus affect subjects' monitoring behavior (cf. Table 9). The factors were Intended Target (/s/ vs. /ʃ/) and Error Magnitude Category (categorical vs. gradient vs. no error). None of the main effects nor the interaction reach significance (Intended Target—$F(1, 23) < 1$; Error Magnitude Category—$F(2, 23) < 1$; interaction—$F(2, 23) = 1.306$, $p = 0.29$). Again, potential perceptual differences between error magnitude categories cannot be due to durational differences.

4.2.2. Results

As for /t - k/, a two-way ANOVA is performed on the d' measure with repeated measures on both factors. Factors are Error Magnitude Category (categorical-gradient-no error)[14] and Intended Target (/s, ʃ/). Again, an intended target is defined as target the speaker was instructed to pronounce in the production experiment. Table 10 shows the mean d' values for the two factors.

For the factor Error Magnitude Category, Mauchly's test of sphericity reaches significance, indicating a violation of sphericity assumptions; the degrees of freedom for that factor have thus

[14] Note that the factor Error Type cannot be divided into four levels as for /t - k/, since for /s -ʃ/, we only take TB into consideration. Thus for /s/, TB errors are intrusion and for /ʃ/ they are reduction errors; there are no intrusion errors for /ʃ/ and vice versa no reduction errors for /s/.

Table 10
Means with standard deviations in parentheses for *d′* results for /s/ and /ʃ/ grouped by error magnitude category

	Intended target	
Error magnitude	/s/	/ʃ/
Categorical	−0.43 (0.19)	−0.98 (0.49)
Gradient	2.99 (0.42)	0.59 (0.32)
No error	3.33 (0.33)	3.78 (0.22)

Table 11
Means with standard deviations in parentheses for reaction time for /s/ and /ʃ/ grouped by error magnitude category

	Intended target	
Error magnitude category	/s/	/ʃ/
Categorical	894.38 (38.43)	1017.6 (66.59)
Gradient	840.42 (35.03)	860.13 (52.49)
No error	782.81 (34.58)	788.59 (27.13)

been adjusted by using Greenhouse–Geisser Epsilon correction. Factor Error Magnitude Category is significant ($F(1.129,\ 12.417) = 172.832$; $p < 0.0001$). Factor Intended Target is not significant ($F(1,\ 11) = 1.96$; $p = 0.189$); the interaction effect reaches significance ($F(2,\ 22) = 50.07$, $p < 0.001$).

A one-way ANOVA on the interaction means with a posthoc test (Ryan–Einot–Gabriel–Welsch Multiple Range Test, $\alpha = 0.01$) shows that the sensitivity values for /ʃ/ are significantly different for categorical, gradient and no error. For /s/, on the other hand, only categorical errors significantly affect sensitivity. There is no statistically significant difference between the no error and gradient error categories. Overall, the perception of /ʃ/ was found to be more affected by error than was the perception of /s/.

The same analysis design is used with reaction time as data points instead of the sensitivity measure. Only hit responses enter into the analysis. Responses are averaged within an error category for a given subject (e.g., gradient reduction /ʃ/, no error /s/, etc.). Only one /s/ token with a full TB error generated no hit responses at all across subjects. Generally, hit rates range from 7% to 100%, with an average of 42% for tokens with categorical errors, 74% for tokens with gradient errors and 95% for error-free tokens.

Table 11 shows the reaction time means and standard deviations for /s/ and /ʃ/ according to error magnitude category. The factor Intended Target is not significant ($F(1,\ 10) = 1.595$; $p = 0.235$). The factor Error Type with the levels gradient–categorical–no error reaches significance ($F(2,\ 20) = 33.266$; $p < 0.0001$). The interaction effect is not significant ($F(2,\ 20) = 2.549$; $p = 0.103$). For both /s/ and /ʃ/ we can see a linear trend with response time increasing from no error to gradient to categorical error.

Table 12
Correlation results for absolute gestural magnitude and reaction time

Tract variable	Intended target	*N*	*r*	*p*
TB	ʃ	144	−0.172	0.039
TB	*s*	144	0.321	<0.0001

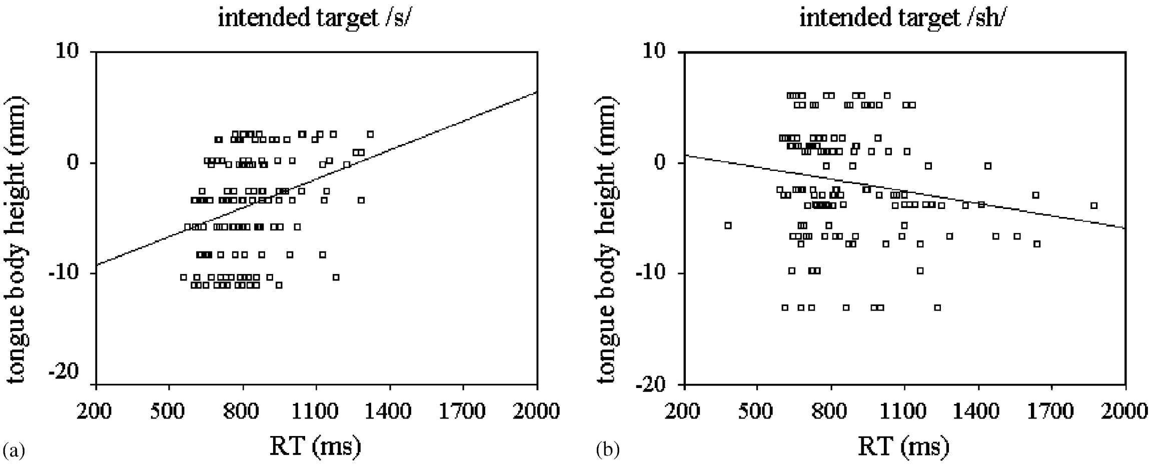

Fig. 3. Scatter plots for correlations of TB magnitude with reaction times. Only hit responses are included. Each data point represents the averaged responses for multiple presentations of the same token within a given subject. (a) Correlation TB height - response time for intended target /s/. (b) Correlation TB height - response time for intended target /ʃ/.

For the correlation between constriction magnitude and reaction time, a negative correlation between TB height and reaction time is predicted for /ʃ/, while a positive correlation is expected for /s/ (the higher TB, the longer response time). As for /t - k/, responses for the same token within a subject are averaged. Averaged this way, we have 58 out of 319 instances (29 tokens × 11 subjects) where subjects scored no hits for a given token. Table 12 shows the results. Fig. 3 shows the corresponding scatter plots.

The result for /ʃ/ approaches significance at the 0.01 level and is significant for /s/, with the correlation coefficient for /ʃ/ being considerably smaller than for /s/. The correlation coefficients reflect the predicted directionality.

For the categorical error category, enough false alarm responses occurred for both /s/ and /ʃ/ (above 60% for each) in order to be able to compare response times for hit and false alarm responses. We can thus ask whether reaction times for tokens that were incorrectly identified as /s/ (or /ʃ/)—that is, tokens where the error was perceptually a substitution—are the same or different from tokens that were correctly identified as /s/ (or /ʃ/). Fig. 4 shows the response times averaged across subjects. The bar graphs compare false alarm responses for tokens with categorical errors (e.g., /s/ with categorical TB errors that were identified as /ʃ/) to tokens that were correctly identified as for example, /ʃ/.

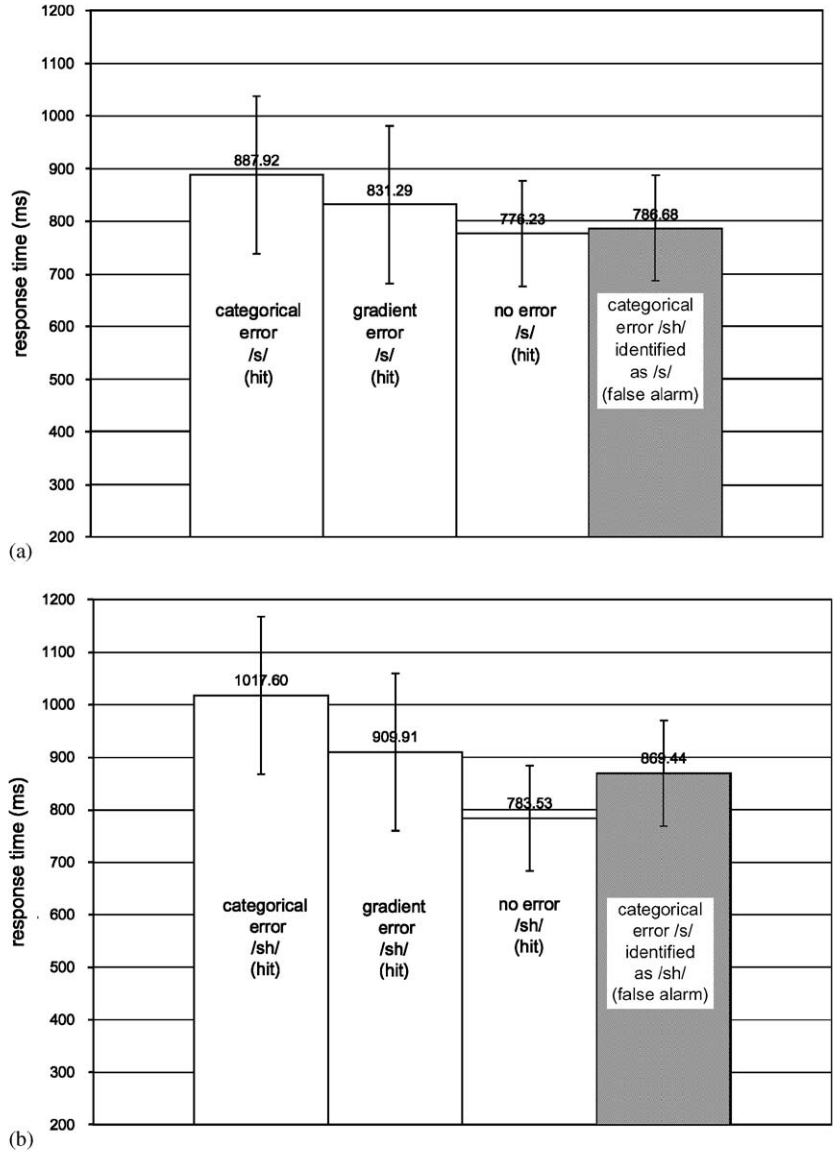

Fig. 4. (a) Reaction time averages for correctly identified /s/ tokens (hit responses, all error categories) vs. /ʃ/ tokens with errors of categorical magnitude that were heard as /s/ (false alarm responses; shaded column). (b) Reaction time averages for correctly identified /ʃ/ tokens (hit responses, all error categories) vs. /s/ tokens with errors of categorical magnitude that were heard as /ʃ/ (false alarm responses; shaded column).

Fig. 4a shows reaction times to tokens that were perceived as /s/. The plain columns show tokens that were intended /s/ by the speaker and heard as /s/ by listeners (referred to as hit responses). The shaded column on the right shows tokens that were intended /ʃ/ by the speaker, but were of categorical errorful magnitude in the articulation, and perceived as /s/ by listeners (referred to as false alarm responses). Listeners' reaction times to these latter tokens with categorical errors in the articulation of /ʃ/ (the shaded, rightmost column) are about as fast as hit responses for error free /s/ tokens. A one-way ANOVA with the categories as displayed in Fig. 4a shows a significant main effect Error Response Type ($F(3,33) = 7.375$, $p = 0.001$). Pairwise comparisons of the mean, using Bonferroni's adjustment for multiple comparisons, reveals that false alarm responses for /ʃ/ are not significantly different from the no error responses for /s/, neither are they significantly different from the hit responses that were recorded for /s/ tokens with full errors ($p = 0.09$).[15] Tentatively, it can be inferred that /ʃ/ tokens with a low TB gesture that were identified as /s/ tend to be perceived as full, unambiguous substitutions.

For Fig. 4b a slightly different picture is apparent. Here, reaction times to tokens that were perceived as /ʃ/ are displayed. The plain columns show tokens that were intended /ʃ/ by the speaker and heard as /ʃ/ by the listener (hit responses). The shaded column on the right shows tokens that were intended /s/ by the speaker, but were of categorical errorful magnitude in the articulation, and perceived as /ʃ/ by listeners. The hit responses for intended /ʃ/ of categorical errorful magnitude are on average slower than the false alarm responses in which a /s/ of categorical errorful magnitude was heard as /ʃ/. These latter categorical error magnitude /s/-tokens that were perceived as /ʃ/ are closest to hit responses for /ʃ/ tokens of the gradient error category. Thus, these false alarm /s/-tokens are heard as more ambiguous than error-free tokens of /ʃ/.[16]

Also in false alarm responses to tokens with categorical errors there emerges an asymmetry between /s/ and /ʃ/: Categorical (as well as gradient) TB errors in the articulation of /s/ lead to a perceptual substitution, that is, /ʃ/ is perceived similarly to error-free /s/. Categorical TB errors in the articulation of /s/ do not lead to an unambiguous /ʃ/-percept. A single factor ANOVA with the levels as displayed in Fig. 4b is significant ($F(3,30) = 9.302$; $p < 0.001$). Pairwise comparisons with Bonferroni adjustment for multiple comparisons show that the difference between the hit responses to no error /ʃ/ and false alarm responses to categorical error /s/ approaches significance at $p = 0.016$. This latter category is not significantly different from any of the other two hit categories.

On the whole, the results for /s - ʃ/ show a slight directionality in that the identification of /ʃ/ is more variable under error as the identification of /s/. However, the asymmetry is overall relatively weak, especially compared to the asymmetry obtained for /t/ and /k/ in Experiment 1. It has to be considered whether the error status of tongue tip affected the outcome of the perceptual results for /s/ and /ʃ/. That the sibilants are perceptually almost equally vulnerable to TB errors might be due to the fact that for errorful tokens the TT gesture was also of errorful magnitude. A further factor that potentially influences the results of the perception experiment is that in contrast to /s/, a gestural specification of /ʃ/ further includes an upper lip (UL) protrusion gesture. For the subject whose kinematic data were used for the perceptual experiment, it was not possible to measure a

[15] Categorical error hit responses for /s/ are significantly different from no error responses for /s/. While false alarm responses were on average about 10 ms slower than the no error responses for /s/, this difference is enough to accept the null hypothesis for the comparison between false alarm responses for /ʃ/ and hit responses for categorical errors on /s/.

[16] It is worth pointing out here that the number of double identifications does not differ considerably for /s/ and /ʃ/, that is, specific tokens were only rarely identified as /s/ as well as as /ʃ/ in the different monitoring conditions. For categorical errors during /s/, in 6% of the cases subjects gave at least one response for /s/ and at least one response for /ʃ/ on the same token. For /ʃ/ tokens with categorical errors, this occurred in 7% of the cases.

difference in UL protrusion for /s/ and /ʃ/ in a way that rendered a reliable basis for a statistical determination of errorful UL behavior. Nevertheless, it cannot be excluded that the presence/absence and magnitude of a UL gesture would interact with perception. The perceptual results for /s/ and /ʃ/ therefore cannot be interpreted in the same way as /t/ and /k/ are. For the stops, the experiment demonstrates the perceptual consequences of co-production of two gestures. For the sibilants, the experimental stimuli are less tightly controlled; the experiment reveals the effect of the occurrence of at least one errorful constriction. The results nonetheless speak to the issue of asymmetrical error distributions, since the lack of independence between TT and TB in the articulatory data shows that the occurrence of single-constriction errors between these two vocal organs is comparatively rare. We can expect this factor to constrain error occurrence outside of the laboratory in a similar fashion.

5. General discussion

The results of this study reveal differences in error perceptibility between different consonants and thus call into question the use of transcription in speech error research. Inferences about the speech production system drawn on the basis of speech errors lose explanatory power if the way the errors are recorded does not necessarily capture the essential properties of these errors. Since in Goldstein et al.'s (submitted) and Pouplier's (2003) studies speech errors have been shown often not to be phonologically well-formed, it is not appropriate to rely solely on transcription for the collection of error data. Also, instrumental acoustic measurements will not always be sufficient if the data are used to draw inferences about the speech production system, since the mapping between articulatory events and measured acoustic outcomes can be ambiguous (e.g., Schroeder, 1967; Atal, Chang, Mathews, & Tukey, 1978; Chen, 2003). Due to coarticulation, not all (errorful or nonerrorful) gestures will have acoustic or perceptual consequences, or at least not the same type of consequences.

In the data presented here, asymmetries have been found in production as well as in perception. However, these asymmetries are different in nature from the ones that have been reported by Stemberger (1991a, b). Goldstein et al. (submitted) and Pouplier (2003) show that both /t/ and /k/ exhibit an intrusion bias; there is no production asymmetry between the two stops. An asymmetry does exist in the form of a gestural addition bias, but this phenomenon affects /t/ and /k/ to the same degree. The present experiments demonstrate that it is in perception that the gestural intrusion bias has different consequences. The perception of coronals is more affected by errors than the perception of dorsals. The anti-frequency effect in substitution errors that has been observed for /t/ and /k/ for error data recorded by means of transcription might thus be due to the transcriber's perceptual biases. Crucially, our data do not require an appeal to coronal underspecification in order to explain the apparent anti-frequency effects in error distributions.

For the sibilant pair /s - ʃ/ on the other hand, the perceptual asymmetry cannot explain the directionality effect found in speech error corpora and experiments. On the basis of the perceptual results for gradient errors a very small directionality effect that might be expected would be /s/ replacing /ʃ/, and not /ʃ/ substituting /s/ (although for categorical error magnitude, /s/ is more affected by error than /ʃ/). Since /ʃ/ is more systematically affected by error than /s/, we would expect /ʃ/ to be more often substituted by /s/. This, however, is the opposite of the asymmetry that has been recorded in speech error research. This suggests that the asymmetries cannot be explained by perceptual biases, they must originate in production. Recall that in production, intrusion errors are more frequent than reduction errors; errors in production are dominated by the intrusion bias. For /ʃ/, there is no intrusion bias in terms of TT, since both TT and TB are hypothesized to be actively controlled for in normal production. Thus, the most common error should be an intrusive TB error on /s/. This means that /s/ will systematically be more affected by errors compared to /ʃ/. The data obtained in the production experiment confirm this prediction (cf. Pouplier, 2003).

It can be concluded that asymmetries can originate in production where the gestural structures of the interacting segments are in a subset relationship to each other. Thus, intrusive TB errors only affect /s/, not /ʃ/. Note that this supports a more limited notion of underspecification assumed in a gestural framework: task specific targets are defined for certain tract variables only, not all articulators have a specified position for every segment (Browman & Goldstein, 1992). Again, there is no need to assume /s/ to be underspecified for place of articulation; it does lack, however, a tongue body constriction gesture. It is to be expected that the palatal bias reported in Stemberger (1991a) as well as Shattuck-Hufnagel and Klatt (1979), in which /t/ turns more often into the affricate /tʃ/ than vice versa, can be explained on the same basis.

At first sight it may seem puzzling that listeners compensate for coarticulation in fluent speech (e.g., Mann, 1980; Mann & Repp, 1980), but seemingly fail to do so in the present experiments. In particular, Browman and Goldstein (1995) have shown on the basis of X-ray microbeam data that coronals can be of smaller magnitude and be substantially overlapped by a flanking consonant—seemingly a similar situation to the error tokens here, suggesting listeners might be familiar with coarticulatory structures as found in the errorful tokens from 'normal' speech. Listeners could then be expected to compensate for these coarticulations in errorful utterances as they do in normal speech. Yet there are two crucial differences between this 'normal' overlap pattern and the error data at hand. First of all, coronals in English reduce in final, not in initial position. In the stimuli presented to subjects here, all critical consonants were word initial. Reduced magnitude thus could not be used by listeners to extract information about prosodic position. Secondly, Browman and Goldstein found a *final* coronal to be overlapped by a following initial stop. Yet again in these stimuli, the coronal was always in initial position. This brings up the point that in fluent speech, coarticulation is not really compensated for, but rather it provides information about preceding or following consonants or vowels. In the data at hand, all utterances were single words with a final /p/, which means that for example, information about a tongue dorsum gesture during a /t/ could not be used to identify any neighboring consonant or vowel. The co-production of two gestures rendered conflicting information as to the identity of the target sound. It might be predicted that if, for a /t/ with an errorful TD intrusion, the phrase *top cop* were presented to subjects (with a nonerrorful *cop* utterance) instead of the *top* in isolation, identification of the /t/ should improve (cf. Surprenant & Goldstein, 1998, for a similar effect with truncated vs. untruncated stimuli), yet still it has to be kept in mind that the errorful co-occurrence of two constrictions is crucially different from normal coarticulatory overlap. In errors, an extra copy of a gesture is added, as for example when in the phrase *top cop* a TD gesture intrudes during the /t/ in *top*, this TD gesture appears in *addition* to the canonical TD gesture during /k/ in *cop*.

A question that remains is what prompts these perceptual asymmetries between coronal and dorsal stops. The relatively high vulnerability of coronals in terms of their acoustic properties is nothing new: Byrd (1992) and Chen (2003) analyze the perception of coronals and find them particularly vulnerable to overlap with other consonant gestures. Increasing overlap between a coronal and, in the case of Byrd (1992), a subsequent bilabial gesture, considerably affected the identification rate of the coronal. The bilabial closure, however, was not obscured to the same degree by an overlapping subsequent coronal gesture. Surprenant and Goldstein (1998) likewise find that /t/ was not completely recoverable by listeners when overlapped by a following /p/. Byrd explains this asymmetry by the greater effect the bilabial closure had on the formant values for the coronal than vice versa. In a computational simulation, Chen (2003) demonstrates that coronals overlapped by labials can be algorithmically recovered as reduced, despite their full gestural magnitude. Labials, on the other hand, were not affected by overlap with a following coronal. She attributed this effect to differences in articulator velocity and gestural timing that presumably leads to differences in how much formant transition information is available to the listener. Due to faster TT movement, a coronal has a short transition time that can be hidden by an overlapping lip gesture.

In terms of perceptual consequences, the speech errors of the type examined here can be thought of as related to the perceptual consequences of temporal gestural overlap observed in fluent speech: The perceptual asymmetry apparent in the context of the errorful co-occurrence of two speech gestures can be viewed as analogous to a more general perceptual coronal–noncoronal asymmetry in English.

6. Conclusions

The experiments presented in this paper provide evidence for systematic asymmetries in the perception of speech production errors. These asymmetries combine with the nature of production errors to provide a potential explanation for the patterns of asymmetries reported in speech error corpora. In production, error patterns are generally dominated by an intrusion bias: it is more likely for an errorful gesture to intrude than it is for a target gesture to be reduced. For different segments, this gestural addition bias bears different consequences. For /t - k/, intruding TD gestures during /t/ have a systematic perceptual effect, whereas for /k/ intruding TT gestures do not significantly affect identification or reaction times. For /s - ʃ/, perceptual biases are not the source of distributional asymmetries. Rather, the addition bias translates into a ‘phoneme bias.’ The most likely error to occur is an intruding TB gesture during /s/; the intrusion bias leaves /ʃ/ unaffected by /s/ since /ʃ/ and /s/ both have TT gestures. A /s/ during which a TB gesture intrudes will result in a /ʃ/-like articulatory structure. Perceptually, there is a slight bias in that /ʃ/ was more affected by error than /s/, yet the asymmetry is relatively weak. That several transcription-based studies (Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1991a) have found a strong palatalization bias again suggests the low force of this perceptual asymmetry. The study shows that the concept of coronal underspecification is not needed to explain asymmetries in speech errors. The special vulnerability of coronals should be seen in the context of other studies on the perceptual weakness of coronals in English.

Acknowledgements

Thank you to Larissa Chen for comments and much help in practical matters. Thank you also to Alexei Kochetov, Carol Fowler, Dani Byrd, Donca Steriade, Ken de Jong, and Stefanie Shattuck-Hufnagel and two anonymous reviewers for insightful comments. Work supported by NIH Grant HD-01994 to Haskins Laboratories.

Appendix A. Stimuli for Experiment 1

(see Table 13).

Table 13

Intended target	TD height (mm)	TT height (mm)	Error status	Errorful constriction	Rate	Vowel	Stress	Phrase position
k	12.75	−10.65	No error	No error	Fast	o	Unstressed	Final
k	12.48	−10.56	No error	No error	Fast	o	Stressed	Final
k	12.81	−11.59	No error	No error	Mid	o	Unstressed	Final
k	14.6	−8.48	No error	No error	Slow	o	Stressed	Final
k	12.53	−11.98	No error	No error	Fast	i	Stressed	Initial
k	13.32	−11.73	No error	No error	Mid	i	Stressed	Initial
k	13.96	−10.52	No error	No error	Mid	i	Stressed	Initial
k	13.02	−9.77	No error	No error	Mid	i	Unstressed	Final
k	14.21	−10.59	No error	No error	Mid	i	Stressed	Final
k	14.04	−8.72	No error	No error	Slow	o	Stressed	Initial
k	8.73	−6.38	Gradient	TD	Slow	o	Unstressed	Final
k	8.58	−10.23	Gradient	TD	Fast	i	Stressed	Initial
k	8.78	−8.9	Gradient	TD	Fast	o	Unstressed	Initial
k	9.83	−9.06	Gradient	TD	Slow	o	Unstressed	Initial
k	10.97	−9.52	Gradient	TD	Slow	o	Unstressed	Initial
k	12.16	−6.34	Gradient	TD	Slow	o	Unstressed	Initial
k	11.82	−1.7	Categorical	TT	Fast	o	Stressed	Initial
k	12.72	−0.46	Categorical	TT	Fast	o	Stressed	Final
k	13.11	−2.94	Categorical	TT	Fast	o	Stressed	Final
k	12.88	−1.6	Categorical	TT	Mid	i	Stressed	Initial
k	13.55	0.49	Categorical	TT	Mid	i	Stressed	Initial
k	11.25	−3.26	Categorical	TT	Fast	i	Unstressed	Final
k	11.08	−2.7	Categorical	TT	Fast	i	Unstressed	Final
k	12.88	−2.36	Categorical	TT	Mid	i	Unstressed	Final
k	11.75	−4.11	Categorical	TT	Mid	o	Unstressed	Initial
k	15.41	−1.87	Categorical	TT	Slow	o	Stressed	Initial
k	13.82	−3.81	Gradient	TT	Fast	o	Unstressed	Final
k	11.88	−4.53	Gradient	TT	Fast	o	Stressed	Final
k	12.91	−4.84	Gradient	TT	Mid	o	Unstressed	Final
k	14.14	−3.29	Gradient	TT	Slow	o	Stressed	Final
k	12.61	−5.51	Gradient	TT	Fast	o	Unstressed	Initial
k	12.88	−1.12	Gradient	TT	Fast	i	Stressed	Initial
k	13.08	−2.66	Gradient	TT	Fast	i	Stressed	Initial

Table 13 (*continued*)

Intended target	TD height (mm)	TT height (mm)	Error status	Errorful constriction	Rate	Vowel	Stress	Phrase position
k	14.4	−5.34	Gradient	TT	Mid	i	Unstressed	Initial
k	13.6	−3.44	Gradient	TT	Mid	i	Unstressed	Initial
k	13.62	−6.4	Gradient	TT	Mid	i	Stressed	Final
t	4.64	−1.02	No error	No error	Slow	o	Unstressed	Initial
t	3.7	−1.68	No error	No error	Fast	o	Stressed	Initial
t	3.24	−0.35	No error	No error	Slow	o	Stressed	Initial
t	0.3	0.41	No error	No error	Mid	i	Unstressed	Final
t	0.54	−0.26	No error	No error	Mid	i	Stressed	Final
t	1.03	−0.47	No error	No error	Mid	i	Stressed	Initial
t	-0.265	−0.6	No error	No error	Fast	i	Stressed	Initial
t	1.09	−0.7	No error	No error	Mid	i	Stressed	Initial
t	3.52	−1.44	No error	No error	Fast	o	Stressed	Final
t	4.41	−1.17	No error	No error	Slow	o	Stressed	Final
t	13.44	−0.69	Categorical	TD	Fast	o	Unstressed	Final
t	11.99	−0.97	Categorical	TD	Fast	o	Unstressed	Initial
t	13.43	−1.42	Categorical	TD	Mid	o	Stressed	Initial
t	6.94	−0.07	Categorical	TD	Slow	o	Unstressed	Initial
t	10.1	0.53	Categorical	TD	Fast	i	Unstressed	Final
t	11.93	−0.7	Categorical	TD	Mid	i	Unstressed	Final
t	13.27	−1.14	Categorical	TD	Mid	i	Stressed	Final
t	9.67	−0.77	Categorical	TD	Fast	i	Stressed	Initial
t	11.42	−1.2	Categorical	TD	Fast	i	Stressed	Initial
t	12.94	−2.08	Categorical	TD	Fast	o	Stressed	Final
t	11.15	1.76	Categorical	TD	Fast	o	Stressed	Initial
t	5.02	−3.18	Gradient	TD	Mid	o	Stressed	Initial
t	9.45	−0.89	Gradient	TD	Fast	o	Unstressed	Final
t	2.69	−0.49	Gradient	TD	Fast	i	Unstressed	Final
t	5.71	−0.49	Gradient	TD	Mid	i	Stressed	Final
t	3.99	−1.13	Gradient	TD	Fast	i	Stressed	Initial
t	8.57	−0.72	Gradient	TD	Fast	o	Stressed	Final
t	7.95	−0.79	Gradient	TD	Fast	i	Stressed	Initial
t	10.69	−1.78	Gradient	TD	Mid	i	Stressed	Initial
t	10.86	−0.88	Gradient	TD	Mid	o	Stressed	Final
t	4	−4.29	Gradient	TT	Mid	o	Stressed	Initial
t	2.94	−4.46	Gradient	TT	Mid	o	Stressed	Initial
t	6.05	−3.11	Gradient	TT	Slow	o	Unstressed	Initial
t	2.41	−2.79	Gradient	TT	Mid	i	Stressed	Initial
t	4.45	−4.27	Gradient	TT	Mid	o	Stressed	Final

Appendix B. Stimuli for Experiment 2

(see Table 14).

Table 14

Intended target	TB height (mm)	Error status of TB	Error status of TT	Rate	Vowel	Stress	Phrase position
ʃ	−7.32	Categorical	Categorical	Mid	o	Unstressed	Initial
ʃ	−3.75	Categorical	Categorical	Slow	o	Stressed	Initial
ʃ	−0.28	Categorical	Gradient	Slow	o	Unstressed	Initial
ʃ	−9.78	Categorical	Categorical	Fast	o	Unstressed	Final
ʃ	−13.1	Categorical	Categorical	Mid	o	Stressed	Final
ʃ	−6.58	Gradient	Categorical	Fast	o	Stressed	Initial
ʃ	−2.91	Gradient	Gradient	Fast	o	Stressed	Initial
ʃ	−3.81	Gradient	Categorical	Mid	o	Unstressed	Initial
ʃ	−5.63	Gradient	Gradient	Fast	o	Stressed	Final
ʃ	−2.39	Gradient	No error	Fast	o	Stressed	Final
ʃ	1.04	No error	No error	Fast	o	Stressed	Initial
ʃ	1.52	No error	No error	Fast	o	Unstressed	Final
ʃ	2.27	No error	No error	Mid	o	Stressed	Final
ʃ	5.23	No error	No error	Slow	o	Unstressed	Final
ʃ	6.09	No error	No error	Slow	o	Stressed	Final
s	2.63	Categorical	Gradient	Slow	o	Stressed	Final
s	6.46	Categorical	Categorical	Slow	o	Stressed	Final
s	0.97	Categorical	Categorical	Fast	o	Stressed	Initial
s	0.2	Categorical	Categorical	Fast	o	Unstressed	Initial
s	−2.56	Categorical	Gradient	Mid	o	Stressed	Initial
s	−3.34	Gradient	Gradient	Fast	o	Unstressed	Final
s	−3.42	Gradient	Gradient	Mid	o	Stressed	Initial
s	−0.11	Gradient	Gradient	Slow	o	Stressed	Initial
s	2.15	Gradient	Gradient	Slow	o	Stressed	Initial
s	−10.3	No error	No error	Fast	o	Unstressed	Final
s	−11.04	No error	No error	Fast	o	Unstressed	Final
s	−5.77	No error	No error	Slow	o	Unstressed	Final
s	−8.33	No error	No error	Mid	o	Unstressed	Initial
s	−5.76	No error	No error	Slow	o	Unstressed	Initial

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.

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. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica: International Journal of Speech Science*, *49*, 155–180.

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In Bell-Berti, F., & Raphael, L. J. (Eds.), *Producing speech: contemporary issues* (pp. 19–33). Woodbury, NY: AIP Press.

Byrd, D. (1992). Perception of assimilation in consonant clusters: A gestural model. *Phonetica*, *49*, 1–24.

Chen, L. (2003). The origins in overlap of place assimilation. In Garding, G., & Tsujimura, M. (Eds.), *WCCFL 22. Proceedings of the XXIIth West Coast Conference on Formal Linguistics* (pp. 137–150). Somerville, MA: Cascadilla Press.

Cohen, A. (1980). Correcting of speech errors in a shadowing task. In Fromkin, V. A. (Ed.), *Errors in linguistic performance. Slips of the tongue, ear, pen, and hand* (pp. 157–163). New York: Academic Press.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, *25*(2), 257–271.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, *1*, 153–156.

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

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

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.

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. (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (submitted) Gestural action units slip in speech production errors.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory. A user's guide*. Cambridge: Cambridge University Press.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics*, *28*(5), 407–412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the ʃ–[s] distinction. *Perception and Psychophysics*, *28*(3), 213–228.

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. (1975). Encoding sensitivities to phonological markedness and transition probability: Evidence from spoonerisms. *Human Communication Research*, *2*, 351–361.

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

Nooteboom, S. G. (1973). The tongue slips into patterns. In Fromkin, V. A. (Ed.), *Speech errors as linguistic evidence* (pp. 144–156). The Hague: Mouton.

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.

Pouplier, M. (2003). *Units of phonological encoding: Empirical evidence*. PhD dissertation, Yale University.

Schroeder, M. R. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of the Acoustical Society of America*, *41*(4), 1002–1010.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In Cooper, W. E., & Walker, E. C. T. (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 MacNeilage, P. F. (Ed.), *The production of speech* (pp. 109–136). New York: Springer.

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition*, *42*, 213–259.

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., & Stoel-Gammon, C. (1991). The underspecification of coronals, evidence from language acquisition and performance errors. In Paradis, C., & Prunet, J. -F. (Eds.), *The special status of coronals. internal and external evidence* (pp. 181–199). San Diego: Academic Press.

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

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

Stemberger, J. P. (1991b). Radical underspecification in language production. *Phonology*, *8*, 73–112.

Surprenant, A. M., & Goldstein, L. (1998). The perception of speech gestures. *Journal of the Acoustical Society of America*, *104*(1), 518–529.

Tent, J., & Clark, J. E. (1980). An experimental investigation into the perception of slips of the tongue. *Journal of Phonetics*, *8*(3), 317–325.

Toothacker, L. E. (1993). *Multiple comparison procedures*. Thousand Oaks, CA: Sage Publications.