Short title: Automatic Analysis of Speech Errors

Automatic Analysis of Slips of the Tongue: Insights into the Cognitive Architecture of Speech Production

Matthew Goldrick[a*], Joseph Keshet[b*], Erin Gustafson[a], Jordana Heller[a], Jeremy Needle[a]

[a]Department of Linguistics, Northwestern University
[b]Department of Computer Science, Bar-Ilan University

*To whom correspondence should be addressed:

Matthew Goldrick
Department of Linguistics, Northwestern University
2016 Sheridan Rd.
Evanston, IL 60208 USA
+1 (847) 491-8053
matt-goldrick@northwestern.edu

Joseph Keshet
Department of Computer Science
Bar-Ilan University
Ramat Gan, 52900 Israel
+972 -3-7384378
joseph.keshet@biu.ac.il

Keywords: Speech production; automatic phonetic analysis; speech errors; machine learning; structured prediction

Abstract

Traces of the cognitive mechanisms underlying speaking can be found within subtle variations in how we pronounce sounds. While speech errors have traditionally been seen as categorical substitutions of one sound for another, acoustic/articulatory analyses show they partially reflect the intended sound. When "pig" is mispronounced as "big," the resulting /b/ sound differs from correct productions of "big," moving towards intended "pig"—revealing the role of graded sound representations in speech production. Investigating the origins of such phenomena requires detailed estimation of speech sound distributions; this has been hampered by reliance on subjective, labor-intensive manual annotation. Computational methods can address these issues by providing for objective, automatic measurements. We develop a novel high-precision computational approach, based on a set of machine learning algorithms, for measurement of elicited speech. The algorithms are trained on existing manually labeled data to detect and locate linguistically relevant acoustic properties with high accuracy. Our approach is robust, is designed to handle mis-productions, and overall matches the performance of expert coders. It allows us to analyze a very large dataset of speech errors (containing far more errors than the total in the existing literature), illuminating properties of speech sound distributions previously impossible to reliably observe. We argue that this provides novel evidence that two sources both contribute to deviations in speech errors: planning processes specifying the targets of articulation and articulatory processes specifying the motor movements that execute this plan. These findings illustrate how a much richer picture of speech provides an opportunity to gain novel insights into language processing.

1. Introduction

The acoustic and articulatory properties of speech vary from moment to moment; if you repeat a word several times, no two instances will be precisely the same. Hidden within this variation are traces of the cognitive processes underlying language production. For example, when repeatedly producing a word, you will tend to slightly reduce its duration—reflecting (in part) the ease of retrieving the word from long term memory (Kahn & Arnold, 2012; Lam & Watson, 2010). Such effects can also be found at the level of individual speech sounds within a word. One such effect can be observed in bilingual speakers' pronunciations of second language speech sounds. Such sounds are more accented when speakers have recently produced a word in their native language, relative to cases where the same speaker has just produced sounds in the second language (Balukas & Koops, 2015; Goldrick, Runnqvist, & Costa, 2014; Olson, 2013). This suggests that the difficulty of retrieving words and sounds when switching languages can modulate how sounds are articulated.

Here, we focus on one source of evidence that has played a key role in theories of language production: speech errors (Fromkin, 1971, et seq.). Errors involving the mis-production of sounds ("pig" mispronounced as "big") reveal the graded influence of intended productions on articulation. Errors simultaneously reflect acoustic/articulatory properties of both the target and error outcome (Frisch & Wright, 2002; Goldrick, Baker, Murphy, & Baese-Berk, 2011; Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; McMillan & Corley, 2010; McMillan, Corley, Lickley, 2009; Pouplier, 2007, 2008). Such effects are consistent with theories of language production incorporating continuous, distributed mental representations in the cognitive process underlying the planning (Dell, 1986; Goldrick & Blumstein, 2006; Smolensky, Goldrick, & Mathis, 2014; Plaut & Shallice, 1993) and articulation of speech sounds

(Goldstein, et al., 2007; Saltzman & Munhall, 1989). According to these theoretical perspectives, articulation reflects subtle, gradient variation in the representational structures and cognitive processes underlying speech (e.g., variation in the degree to which the native language is activated can yield graded changes in the degree of accent in non-native speech; partial activation of target sounds can influence how errors are articulated).

While studies of phonetic variation have provided a rich source of information about language processing, most researchers have relied on manual annotation to obtain accurate data. This approach suffers from two critical flaws. It is highly resource intensive; a single experiment in our lab (Goldrick et al., 2011) required over 3,000 person-hours for analysis. With respect to speech error studies (as discussed below), this has prevented researchers from obtaining the data required to reliably evaluate different hypotheses. Second, this approach is fundamentally subjective: manual labels reflect the judgments of annotators. This presents a barrier to replication.

Recent studies have aimed to address these issues through computational methods that automatically measure acoustic properties of speech (e.g., Gahl, Yao, & Johnson, 2012; Labov, Rosenfelder, & Freuhwald, 2013; Yuan & Liberman, 2014). These methods eliminate subjective judgments while enormously reducing the resources required for analysis. Although this has provided great advances in studies of phonetic variation, existing methods do not provide a comprehensive solution. They have not provided the fine granularity of measurement necessary to reliably measure differences at the level of individual speech sounds (specifically, consonant sounds). Furthermore, these existing methods require a complete transcription of the observed speech prior to phonetic analysis. This is a major burden, particularly for paradigms that are designed to produce tremendous variation in production (e.g., speech errors).

In this work, we propose a novel computational framework for automatic analysis of speech appropriate for evaluating hypotheses relating to the phonetics of speech errors. This is based on a set of algorithms in machine learning (Keshet, Shalev-Schwartz, Singer, & Chazan, 2007; McAllester, Hazan, & Keshet, 2010; Sonderegger & Keshet, 2012). Our automatic approach matches the performance of expert manual coders and outperforms algorithms used in the existing psycholinguistic literature. The analyses reveal novel properties of the phonetics of speech errors. Furthermore, we show (via a power analysis) that reliable investigation of the properties of individual speech sounds requires datasets larger than those used in previous work. These findings show how automatic analysis creates an opportunity to gain a much richer, objective, and replicable picture of acoustic variation in speech.

1.1 Phonetic Variation in Sound Substitution Errors

One key source of evidence for the structure of the cognitive mechanisms underlying language production is speech errors (Fromkin, 1971). Sound substitution errors (e.g., intending to say *bet*, but producing *pet*; written as *bet*→*pet*) have been studied in the laboratory by asking participants to rapidly produce artificial tongue twisters composed of syllables with alternating contrasting sounds (*pet bet bet pet*; Wilshire, 1999). Based on transcriptions of speech, it was long assumed that such errors reflect the categorical substitution of one sound for another (Dell, 1986; Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979). However, more recent quantitative analyses of the phonetic (acoustic/articulatory) properties of errors have revealed that errors systematically differ from corresponding correct productions—a deviation that reflects properties of the intended sound (Frisch & Wright, 2002; Goldrick et al., 2011; Goldrick & Blumstein, 2006; Goldstein, et al., 2007; McMillan & Corley, 2010; McMillan, Corley, Lickley, 2009;

Pouplier, 2007, 2008). For example, an important acoustic cue to the distinction between words like *pet* and *bet* is voice onset time (VOT), the time between the release of airflow (e.g., opening the lips) and the onset of periodic vibration of the vocal folds (Lisker & Abramson, 1964). In English, voiceless sounds like /p/ have relatively long VOTs whereas voiced sounds like /b/ have short VOTs (Lisker & Abramson, 1964). In a *bet*→*pet* error, the resulting /p/ sound is distinct from correct productions of the same sound (*pet*→*pet*). The error /p/ tends to have a shorter VOT—which makes it more similar to the intended sound /b/. The complementary pattern is found for errors like *pet*→*bet*; the error /b/ tends to have a longer VOT than the corresponding sound in *bet*→*bet*. Note that similar effects are found in non-errorful speech when a competitor word is explicitly primed (e.g., priming *top* while reading the word *cop* aloud yields a blend of /t/ and /k/ articulations; Yuen, Davis, Brysbaert, & Rastle, 2010)

These deviations have been attributed to one of two distinct types of cognitive processes that underlie the production of speech: (i) *planning processes* that construct a relatively abstract specification of the targets of articulation; or (ii) *articulatory processes* that specify the specific motor movements that execute this plan. To illustrate this division, when producing *pet*, planning processes might specify that the initial sound is /p/ but not the precise timing of the associated lip movements; these would be specified during articulatory processing. Below, we outline how different theories have proposed that deviations of errors from correct productions arise at each level of processing.

Within planning processes, many theories of speech production assume that representations are patterns of activation over simple processing units (Dell, 1986). For example, the contrast between *big* and *pig* is represented by graded patterns of activation over units representing speech segments /p/ and /b/. While this type of representation can express arbitrarily

varying combinations of /p/ and /b/, theories typically incorporate mechanisms that constrain the patterns of activation. These mechanisms force planning processes to select relatively discrete representations for production (e.g., primarily activating /p/, with little activation of /b/). A variety of mechanisms have been proposed to account for this, including: boosting activation of one representation relative to alternatives (e.g., Dell, 1986); lateral inhibition that reduces the activation of alternative representations (see Dell & O'Seaghdha, 1994, for a review); and attractors over distributed representations (e.g., Goldrick & Chu, 2014; Plaut & Shallice, 1993; Smolensky et al., 2014). However, these constraints on activation are typically not categorical; while one unit may be highly active, others may remain partially active. This has been proposed as one possible mechanism for producing deviations in speech errors. If the specification of the intended target sound remains partially active, the phonetic properties of the error could be distorted towards the intended target (Goldrick & Blumstein, 2006; Goldrick & Chu, 2014; Smolensky et al., 2014). For example, in *bet*→*pet*, the speech plan could specify the target is 0.9 /p/ and 0.1 /b/—resulting in articulations that combine properties of both sounds.

Articulatory processes could provide an additional source of distortions in speech errors. Such processes specify the continuous, coordinated dynamics of articulator movements that execute the speech plan (Saltzman & Munhall, 1989). Tongue twisters require speakers to rhythmically alternate different configurations of speech gestures (e.g., altering the relative timing of lip opening and glottal movement for /p/ vs. /b/). Research across a variety of domains of action has suggested that alternating different movements is inherently less dynamically stable than repeating synchronous actions. When participants are asked to perform alternating movements under varying response speeds, they spontaneously shift from successful alternation to synchronized movements at fast rates (Haken, Peper, Beek, & DaVertshofer, 1996). If speech

errors in tongue twisters reflected, in part, a similar process—a destabilization of articulation of alternating movements under fast rates—we might expect a similar pattern to emerge. The synchronous production of previously alternating sounds would manifest as a blend of properties of the error and the intended target, providing a second possible mechanism for producing deviations in speech errors (Goldstein et al., 2007, Pouplier, 2007).

Evaluating these two approaches to deviations in errors has been hampered by the relative paucity of phonetic data. For example, studies arguing for an articulatory locus of deviations have often induced errors using repeating sequences (*pet bet pet bet*; e.g., Goldstein et al., 2007). In contrast, studies arguing for a planning locus have often used twisters where the order of pairs of syllables switches within a twister (*pet bet bet pet*; e.g., Goldrick & Blumstein, 2006). Transcription studies suggest that the difference between these two twister types exerts a significant influence on processing (Croot, Au, & Harper, 2010; Wilshire, 1999). Across studies, relative to twisters using repeated syllables sequences, twisters that switch the order of syllables result in higher errors at points where the order of syllable switches (i.e., the first and third positions in a sequence; *pet bet bet pet*). While multiple transcription studies have examined this issue, phonetic studies have not. This likely reflects the high cost of analyzing phonetic data; comparison of syllable orders within items and participants requires collecting twice the amount of data as any single paradigm. The consequence of this methodological divergence has, as of yet, been unexamined. Developing a more efficient means of gathering phonetic data could allow us to bridge results across these two types of studies.

The paucity of phonetic data has also constrained the types of measures that can be examined. While many studies have examined shifts in the mean properties of errors vs. correct productions (e.g., the typical size of deviations away from canonical /b/, towards the intended

/p/), the processing conditions that give rise to speech errors might also influence other distributional properties of productions—in particular, errors might exhibit a different degree of variability than correct productions. The difficulty of production processing may influence phonetic variability. For example, Heisler, Goffman, and Younger (2010) found that children produced novel sound sequences with higher articulatory variability when the strings were not paired with a lexical referent. If participants learned that the sequence was the label for an object, articulation became less variable. However, previous work has not examined whether the processing difficulties that give rise to speech errors in adults might also influence the variability of articulation. This likely reflects the high cost of analyzing phonetic data; the proper assessment of the variability of errors relative to correct productions[1] requires a large number of observations from a substantial number of participants. Each participant must produce a significant number of observations within each condition in order for us to reliably assess the distributional properties of their errors and correct productions. Then, to assess whether such distributional properties are reliably different, we must compare measures across a number of participants. Analyses of this type therefore require a substantial decrease in the cost of gathering phonetic data.

[1] Although the variability of correct vs. error productions has not been contrasted, note that previous work has examined overall variability of productions across conditions (McMillan et al., 2009; McMillan & Corley, 2010).

1.2 Automatic Analysis of Speech

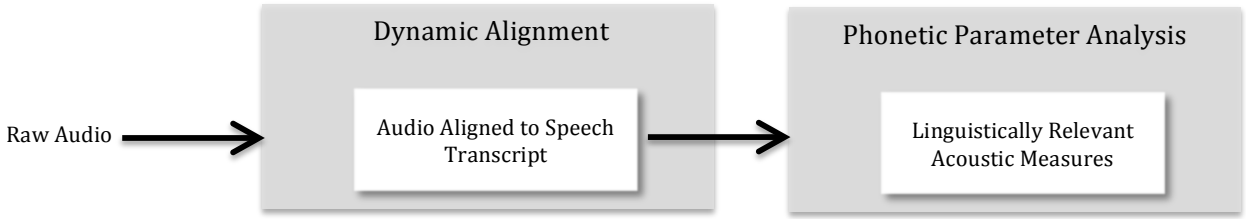


Figure 1. Schematic for automatic analysis of elicited speech. Recorded speech is aligned to a transcript, dynamically selected from a set of possible transcripts (allowing for deviations from the target production—mispronunciations, omitted syllables). These alignments are used to determine windows within which linguistically relevant acoustic properties are measured.

To address the problems associated with limited amounts of speech data, we propose a new approach to the automatic analysis of speech error data. We formulate the general problem as speech sound measurement in studies where speech is elicited by a prompt specifying a target utterance. The objective is to take recordings of such utterances and output an accurate measurement of specific, linguistically relevant dimensions of the acoustic signal (phonetic parameters). As outlined in Figure 1, we approach this problem by first identifying the relevant regions of the acoustic signal for phonetic analysis, and then automatically measuring some linguistically relevant acoustic properties (Section 2 provides full implementation details).

Each of the boxes in Figure 1 corresponds to a learning algorithm that was specially designed to solve the task of phonetic parameter measurement and to minimize the error in the algorithm's prediction for these measures. Unlike problems of classification or regression where the input is a fixed length feature vector and the output is a single bit (such as "grammatical" vs. "not grammatical") or a real number (such as a pitch value), respectively, the input to each of the tasks represented by the boxes in Figure 1 is a structured object (e.g., a variable length acoustic signal), as is the output (e.g., an alignment between phoneme sequences and regions of the acoustic signal; phonetic parameter measurement for particular regions of the acoustic signal).

Structured prediction refers to machine learning models that predict relational information that has structure such as being composed of multiple interrelated parts. A structured prediction algorithm maps the input object along with the target output object into a feature vector space. The algorithm distills to a classifier in this vector space, which is trained to predict the target output object. The classifiers used in this work are based on the large margin concept, meaning that they are trained to separate the target output object from all other output objects with some confidence called *margin*. This allows the trained model to account for perturbations in the vectors in feature vector space due to noise in the speech signal (Crammer, Dekel, Keshet, Shalev-Shwartz, & Singer, 2006). The classifier's confidence can be used to identify noisy input to the classifier or poor classification results, as detailed in Section 2.4.

In contrast to the standard approaches that have been developed for binary classification, each structured prediction task is distinctive: it has a unique evaluation metric, its own set of feature functions, and in many cases requires a non-standard procedure for predicting the target object from an input object, given a set of trained parameters. An overview of our approach is provided in Figure 1. The first box in Figure 1 is a structured prediction algorithm (Keshet et al., 2007, McAllester et al., 2010) that automatically aligns the transcription of the utterance at the level of individual speech sounds (phonemes) with the corresponding portions of the recorded acoustic signal. In contrast to standard existing approaches (Gahl et al., 2012; Labov et al., 2013; Yuan & Liberman, 2014), this transcription is dynamically generated: the target utterance is used to generate several possible transcriptions, allowing for deletion or addition of syllables and mispronunciation of key segments. This eliminates a substantial manual step required by previous approaches. The transcription that best aligns with the recorded acoustics is then used to determine analysis windows for measurement of phonetic parameters. Our state-of-the-art

phoneme alignment algorithm has two advantages relative to existing approaches (Brugnara, Falavigna, & Omologo, 1993, Hosom, 2009): it was designed to minimize the predicted error in the alignment (McAllester et al., 2010); and it extends the representation of the speech acoustics so as to capture temporal regularities in the signal which correlate highly with phoneme boundaries (Keshet et al., 2007). This allows the algorithm to achieve significantly higher accuracy than competing approaches on a standard benchmark (see Section 5 of McAllester et al., 2010).

As noted above, an important acoustic cue that we focus on in our analysis of speech errors is VOT. The second box in Figure 1 is a structured prediction algorithm for measurement of VOT (Sonderegger & Keshet, 2012; see Ryant, Yuan, & Liberman, 2013, for an alternative approach). Many standard approaches measure parameters based on pre-programmed rules developed in consultation with expert annotators (Boyce, Fell, MacAuslan, & Wilde, 2010; Hansen, Gray, & Kim, 2010; Prathosh, Ramakrishnan, & Ananthapadmanabha, 2014; Stouten & van Hamme, 2009). In contrast, the algorithm utilized here was designed to minimize the error in the predicted measurement and had a unique feature set. This novel and unique feature set was designed to represent the acoustic signal with a time resolution of 1 millisecond (based on processing window of 5 milliseconds) as opposed to the 10 millisecond resolution (reflecting a window of 20-25 milliseconds) of standard set feature sets used in automatic speech recognition. By allowing us to measure rapidly changing, short-duration acoustic features, this feature set reflects properties relating to the critical phonetic parameters of our analysis (e.g., those associated with consonant contrasts). Previous research has shown this algorithm can achieve high accuracy, near that of human inter-annotator reliability. For example, using a VOT dataset collected in our laboratory, the algorithm's measurements had correlation $r = 0.992$ with human

annotations, compared with $r = 0.987$ for two human annotators (Sonderegger & Keshet, 2012). Comparison of the VOT algorithm used in our experiments to most of the available automatic methods for four different benchmarks is detailed in Section VII of Sonderegger and Keshet (2012).

This yields acoustic data from speech recordings without requiring human intervention at any intermediate analysis steps. Software implementing each stage of processing is publicly available (https://github.com/jkeshet/tongue_twisters), allowing any laboratory to replicate the analysis procedures on novel data. We used this approach to examine—in a single experiment—the VOT of over 68,000 syllables. In comparison, the amount of data examined across all existing studies (through 2011) is less than 43,000 syllables in total (Frisch & Wright, 2002; Goldrick et al., 2011; Goldrick & Blumstein, 2006; Goldstein, et al., 2007; McMillan & Corley, 2010; McMillan, et al., 2009; Pouplier, 2003: Experiment 2; 2007, 2008). This amount of data allowed us to examine two issues unaddressed in previous work: whether the distinct types of tongue twisters utilized in previous work yield distinct phonetic effects in speech errors; and whether speech errors exhibit differences in variability as well as mean phonetic properties relative to correct productions.

2. Materials and Methods

2.1 Participants

Thirty-four native English speakers (21 women) from the Northwestern University community participated. These individuals reported no history of speech or language impairment. They received financial compensation or course credit.

2.2 Materials

Tongue twisters were composed of syllables with initial consonants contrasting in voicing (e.g., *post-boast*). Forty-eight pairs of syllables were selected, evenly distributed across labial (/p/, /b/), alveolar (/t/, /d/) and velar (/k/, /g/) place of articulation. For each syllable, four tongue twisters were generated, crossing syllable order (switching, ABBA vs. repeating, ABAB) and which member of the pair was placed first (e.g., within ABBA, *post boast boast post* vs. *boast post post boast*). The 48 pairs were generated from 24 quadruplets of syllables (a full list of twisters is provided in the appendix). These consisted of a pair of words (*boast-post*) matched with a pair consisting of a word and nonword (*bolt-polt*)[2].

Syllable order was blocked, such that each participant saw all of the tongue twisters in one order, and then the same tongue twisters in the other order. Block order was counterbalanced across participants. Due to a software error, the last trial was omitted for participants 2-4.

2.3 Procedure

Each target sequence was presented to participants on a computer screen in a sound-attenuated room. Productions were recorded using a head-mounted microphone. Participants practiced each tongue twister once slowly (1 syllable/second) and then repeated it three times quickly (2.5 syllables/second) in time to a metronome. Only tokens from the fast repetitions of each sequence were analyzed. Trial onset and the onset of fast repetitions were self-paced.

[2] This lexicality manipulation followed that of previous work (Goldrick & Blumstein, 2006, Frisch & Wright, 2002, McMillan et al., 2009). As discussed in the Supplementary Materials, this did not consistently influence error rates or phonetic properties of errors, consistent with results suggesting effects of lexicality vary across experimental conditions (Dell, 1986, Hartsuiker, Corley, & Martensen, 2005).

2.4 Automated Acoustic Analysis

Figure 2 elaborates Figure 1, providing a more detailed overview of our approach to analysis of speech. In this study, a recording consisted of the 3 fast repetitions of a tongue twister sequence.

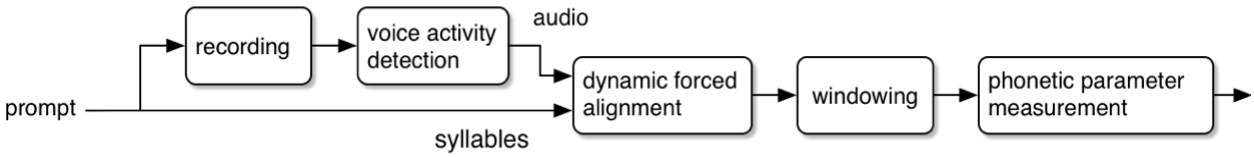


Figure 2. Detailed schematic for automatic analysis of elicited speech. In such tasks, participants are provided with a prompt that specifies the sequence of syllables making up a target utterance. Recorded speech is separated from surrounding silence by detecting voice activity. The speech and individual target syllables are then aligned to one another, dynamically allowing for deviations from the target production (mispronunciations, omitted syllables). These alignments are used to determine windows within which linguistically relevant acoustic properties are measured.

For the current study, each of these steps was implemented as follows (analysis code for these and all subsequent sections is available at https://github.com/jkeshet/tongue_twisters):

- **Voice activity detection (VAD)**. As a preprocessing step, portions of silence greater than 200 milliseconds were removed from the acoustic signal. This was done using a Passive-Aggressive binary classifier (Crammer et al., 2006) that was trained on a set of 69 utterances from 8 participants; these were annotated for speech vs. non-speech portions. The speech signal was framed into 10 millisecond frames, and the first 8 mel-frequency cepstral coefficients (MFCCs) plus an energy coefficient were extracted from each frame (27 features). The input to the classifier was the acoustic features along with the features from the previous two frames and the features from the following two frames (overall 27 x 5 = 135 features). The classifier was trained with a radial basis function kernel ($\sigma = 2.6$, $C = 1.0$), and attained frame-level accuracy of 92% on a 10-fold cross validation. We smoothed the frame-level

predictions by a median filter of size 100. This resulted in a set of smoothed intervals of speech. From those intervals we picked the longest interval as our main processing portion. Overall the VAD algorithm always detected the main interval of the speech, as there was no background noise. In some rare cases when there were non-speech portions (laughing, noises produced by touching the mic, etc.) within the main processing interval, they were detected by checking the confidence of the forced aligner and the deviation of the syllables from the mean as described in the next bullet.

- **Dynamic forced alignment**. We used a structured prediction algorithm developed in Keshet et al. (2007) and McAllester et al. (2010). This was trained on the TIMIT corpus (Garofolo et al., 1993), consisting of 5.4 hours of clean read speech, transcribed at the phoneme level. The input to the aligner was the speech-specific portion of the audio and a transcript reflecting the syllables specified in the *target* tongue twister. Note that because the forced aligner tracks the full range of possible phonemes, it is able to handle errors in the productions. For example, it can align /b oʊ l/ when the actual production was /p oʊ l/, as the target and error phonemes are highly similar. Alignments were computed for a range of transcripts, varying the number of syllables from 7 to 14 (where the target number of syllables is 12). The actual number of syllables was chosen to be the one which resulted with the highest confidence of the forced aligner. Given that productions were intended to be produced at regular intervals (as specified by the metronome), we selected the transcript where the duration of the interval between each initial consonant exhibited the least amount of variation. We did that by computing the average squared

difference between the centers of two adjacent syllables. Based on inspection of the aligner's output, thresholds were set to exclude poor alignment fits; this resulted in the exclusion of 5.2% of the total possible target productions (N = 4104 / 78,300). Syllables beyond the 12 target repetitions were excluded. Based on the selected transcripts, the algorithm estimated that 232 syllables were omitted by participants (0.3% of total possible target productions).

- **Windowing**. Using this transcript, processing windows for the phonetic parameter measurement algorithm were defined. This included 20% of the preceding segment and 100 milliseconds into the following segment, where the reference point was the detected start of the burst. If the extended window boundaries resulted in overlapping analysis windows for adjacent syllables or exceeded the boundaries of the file, we trimmed them to the longest available option.
- **Phonetic parameter measurement**. The algorithm developed in Sonderegger and Keshet (2012) was trained on a set of over 19,000 syllables manually coded in a previous study (Goldrick et al., 2011). The acoustic signal within each analysis window was represented by a set of acoustic features reflecting the onset and offset of VOT. The resulting classifier was used to estimate VOT within each of the analysis windows identified above. Inspection of initial algorithm performance revealed that estimated VOTs of less than 5 milliseconds were typically errors; these observations were therefore excluded (N = 5474, 7.4% of the 73,964 initial consonants present in the selected transcripts). Note that the algorithm does not detect prevoicing; however, with this population, under tongue twister production conditions, prevoicing is very rare (Goldrick et al., 2011). The upper 0.5% of remaining observations (VOTs

exceeding 144 milliseconds) were excluded as outliers. This yielded a total of 68,159 VOTs for analysis.

2.4.1 Categorization of Productions as Voiced vs. Voiceless

English VOT distributions are bimodal, similar to that of many other languages distinguishing two voicing categories in initial position; this empirical pattern has long been assumed to reflect the presence of two distinct planning representations reflecting voiced vs. voiceless (Lisker & Abramson, 1964). The bimodal distributions arise because one planning representation nearly completely dominates the characteristics of each articulation (see the discussion of "nearly discrete" selection mechanisms in §1.1). This generative model can be approximated by a discrete mixture model, where each production is assumed to arise from one of two distinct components[3].

Using the R package *mixdist* (MacDonald, 2012), a mixture of two gamma distributions (representing the distinction between voiced vs. voiceless sound categories) was fit to the VOTs of each participant at each place of articulation. This estimates the mean and variance of two gamma distributions as well as their relative contribution to the overall VOT distribution. Gamma distributions were utilized instead of normal distributions as they provided a better fit to the long right-tailed VOT distributions (Goldrick et al., 2011).

Two mixture model fits at each of two initial parameter settings were calculated (varying the location of the mean of the voiced component); the model with the best fit to the data was utilized. These model fits provided a maximum likelihood threshold, allowing us to classify each

[3] A crucial area for extension of this work would be to move beyond this approximation to explicitly model the *quasi*-discreteness of selection, including gradient co-activation in the generative model.

production as voiced or voiceless. Errors were defined as cases where the intended voicing did not match the voicing of the production[4].

2.4.2. Linking of syllables to target transcript

When the transcript identified by our algorithm contained fewer than 12 syllables, the alignment of these *N* syllables to the 12 target syllables had to be computed (so that it was clear which syllables had been omitted). This was determined by minimizing the edit distance between the target and observed sequences of voiced vs. voiceless initial consonants (preferring deletions to replacements, barring insertions into the target string). In the case of multiple such alignments, those in which deletions occurred at the end of the string were preferred.

3. Results

Across participants, the mean overall accuracy was 89.9% (estimated[5] 95% confidence interval [87.7%, 91.8%]). As has been observed in previous studies, there was considerable variation across individuals (range: 74.3% - 97.2%). Table 1 provides a breakdown of accuracy by place of articulation and voicing.

[4] For one participant, high variance in the voiced component of the mixture model caused a long VOT (> 120 msec) to be classified as voiced. This was excluded from the analysis.
[5] For these descriptive statistics, confidence intervals via a bootstrap procedure over participant means with 1,000 replicates.

Table 1. Mean (across participants) proportion correct on initial consonants, separated by place of articulation and voicing (estimated 95% confidence interval in brackets).

	Alveolar	Labial	Velar
Voiced	89.9% [87.0%, 92.4%]	81.5% [73.8%, 88.0%]	89.1% [83.8%, 92.5%]
Voiceless	91.9% [90.0%, 93.6%]	94.0% [91.6%, 95.8%]	91.2% [89.1%, 93.0%]

We first establish that the automatic approach replicates standard results. We then examine the effects of increasing planning demands; finally, we examine the variability of errors. All of the acoustic data with automatic annotations are available in the Online Speech/Corpora Archive and Analysis Resource (https://oscaar.ci.northwestern.edu/).

3.1 Statistical analysis method

To determine whether the distributional properties (mean, variance) of the phonetic properties of error productions deviated from correct trials, a Monte Carlo method was used to estimate the expected values of these distributional properties in a sample of correct productions equal in size to the set of error productions. We utilized this method as it permitted parallel analyses of changes to means and variance across conditions, allowing us to assess whether such changes were reliable across speakers and items (quadruplets).

Within each speaker, for all the errorful productions on a given syllable pair in a particular twister order (*bet*→*pet* in a repeating twister), the set of corresponding correct productions on the same pair in the same condition (e.g., the repeat order) were selected (*pet*→*pet*). A random sample, equal in size to the number of errors, was drawn from this set of matched correct productions. The distributional properties (mean, standard deviation over participants or items) of this random sample were calculated. This process was repeated 1,000

times to provide an estimate of the chance distribution of these distributional properties in a sample of correct productions of this size.

Given that the sample of correct productions needed to be at least as large as the corresponding set of errors, cases where the number of errors exceeded the number of matched correct productions were excluded. For errors resulting in voiceless consonants, 73 errors (1.8% of 4041 total errors) were excluded. For voiced consonants, 122 errors (4.6% of 2679 total errors) were excluded.

3.2 Validation

We validated the algorithm by comparing the results to three previous results. As shown in Table 2, we replicate previous transcription studies that show higher accuracy in first and third positions in a twister in Repeat vs. Switch orders (Croot et al., 2010; Wilshire, 1999).

Table 2. Mean (across participants) proportion correct on each position in a twister, with difference across orders (estimated 95% confidence interval for difference in brackets). **Bold** indicates significant differences.

Order	Syllable 1	Syllable 2	Syllable 3	Syllable 4
Repeat	91.8%	91.4%	91.9%	87.6%
Switch	88.7%	90.7%	87.8%	89.2%
Repeat–Switch	**3.2%** [1.4%, 5.1%]	0.7% [–0.9%, 2.2%]	**4.1%** [1.7%, 6.7%]	–1.6% [–3.6%, 0.2%]

Second, in non-errorful productions, we replicated the standard pattern of variation in VOT as a function of the place in the oral cavity where airflow was restricted (Cho & Ladefoged, 1999; labial /p/, mean VOT 61.6 milliseconds; tongue tip /t/, 66.8 milliseconds; tongue body /k/ 72.3 milliseconds).

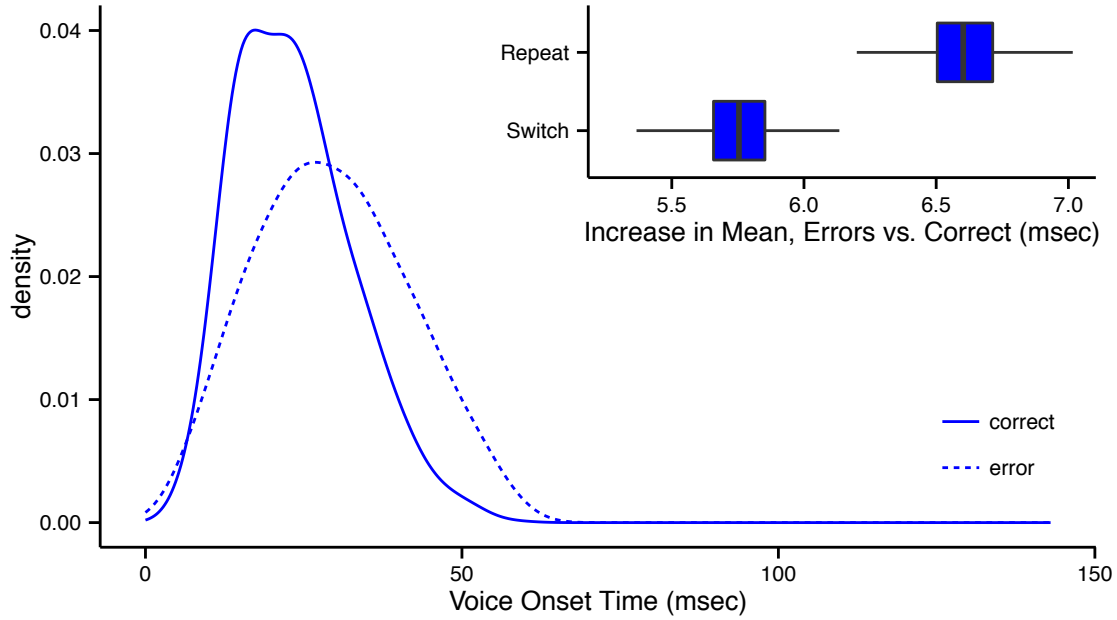
Finally, in errorful productions, we replicated previous work (Goldrick et al., 2011; Goldrick & Blumstein, 2006) showing that errors on voiced and voiceless stops reflect phonetic properties of the intended sound, that is, longer VOTs than correct productions for errors like *pet→bet* (Figure 3A), shorter VOTs for errors like *bet→pet* (Figure 3B). A Monte Carlo method estimated the confidence interval for the difference in means for errors vs. correct productions. As shown in Table 3 and Figure 3, for both voiced and voiceless consonants, this difference was consistently reliable ($p < .0001$ for all comparisons). Similar results were found in a by-items analysis (95% confidence interval for difference from correct productions, voiced repeat: [5.8, 6.4]; voiced switch: [4.9, 5.5]; voiceless repeat: [–17.0, –15.9]; voiceless switch [–12.6, –11.5]; $ps < .0001$).

3.3 Shifts in mean VOT across conditions

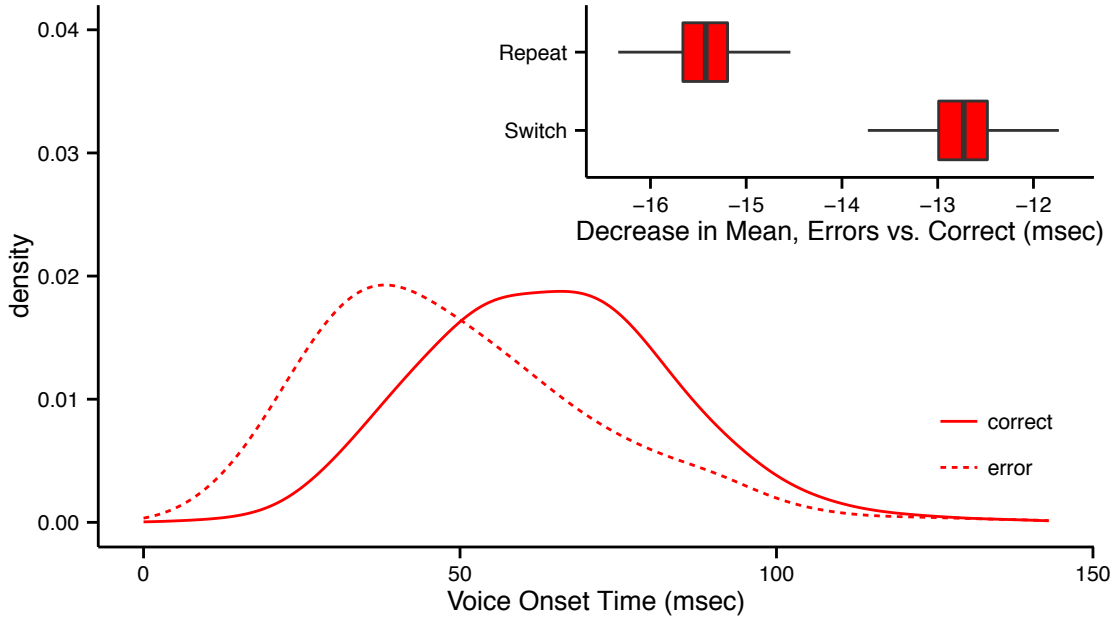
As shown in Figure 3 and Table 3, the switching condition more closely approximated categorical substitutions. That is, errors in this condition consistently exhibited smaller deviations from correct productions of the error outcome than the repeat condition (95% confidence interval for difference in by-participant means across conditions, voiced consonants: [0.5,1.3] milliseconds, $p < .0001$; voiceless consonants: [–3.7, –1.6] milliseconds, $p < .0001$). Similar results were found for by-item means (95% confidence interval, voiced: [0.4, 1.3] milliseconds, $p < .0001$; voiceless: [–5.2, –3.6] milliseconds, $p < .0001$).

Table 3. Mean VOT (across participants) of errors in each tongue twister syllable ordering condition (95% confidence intervals for difference from correct productions in brackets).

	Voiced	Voiceless
Repeat	29.5 [6.3, 6.9]	52.5 [–16.1, –14.8]
Switch	28.0 [5.5, 6.0]	53.1 [–13.6, –12.0]



A.



B.

Figure 3. Shifts in mean of distribution for errors vs. correct productions (A: voiced consonants; B: voiceless consonants). Smoothed density plots characterize the distribution of VOTs for all errors vs. a single random sample of correct productions produced by the same speaker on the same trials. Inset boxplots show the distribution of differences between the mean VOT of errors (across participants) and the mean VOT of 1,000 random samples of matched correct productions.

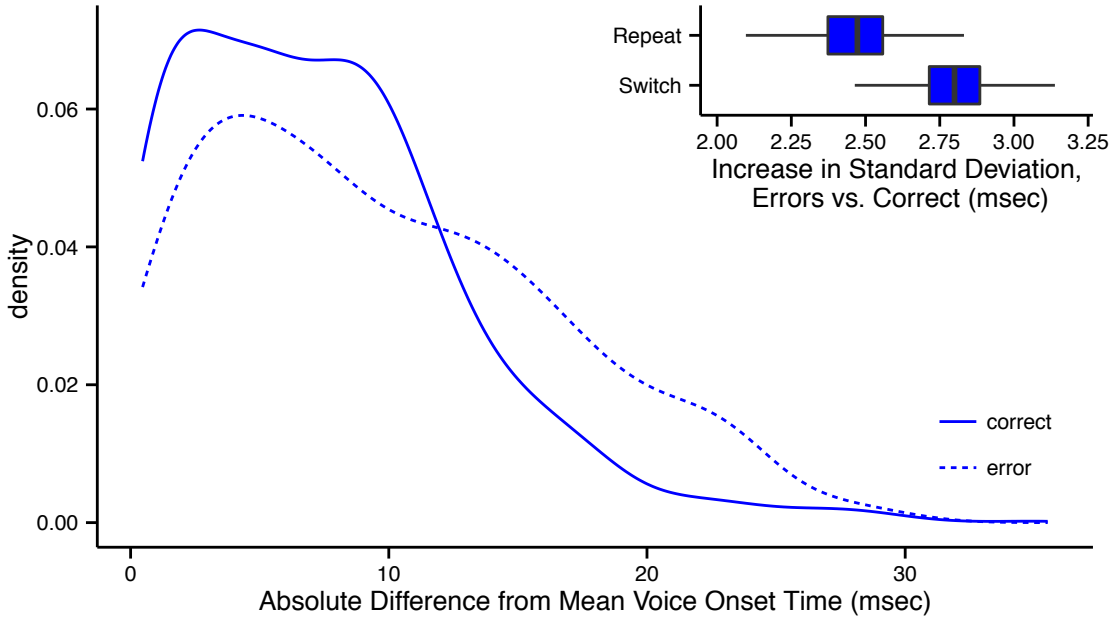
3.4 Phonetic variability of errors

We also find that the variability of error tokens was significantly higher than that of correct productions. Using the same Monte Carlo method, we estimated the 95% confidence interval for the difference in the mean standard deviation for errors vs. correct productions. As shown in Table 4 and Figure 4, we find that across participants this difference was consistently greater than 0 ($p < .0001$). Similar results were found in a by-items analysis (95% confidence interval for difference from correct productions, voiced repeat: [3.0, 3.6]; voiced switch: [2.1, 2.7]; voiceless repeat: [1.4, 2.5]; voiceless switch: [2.1, 3.2]; $ps < .0001$)

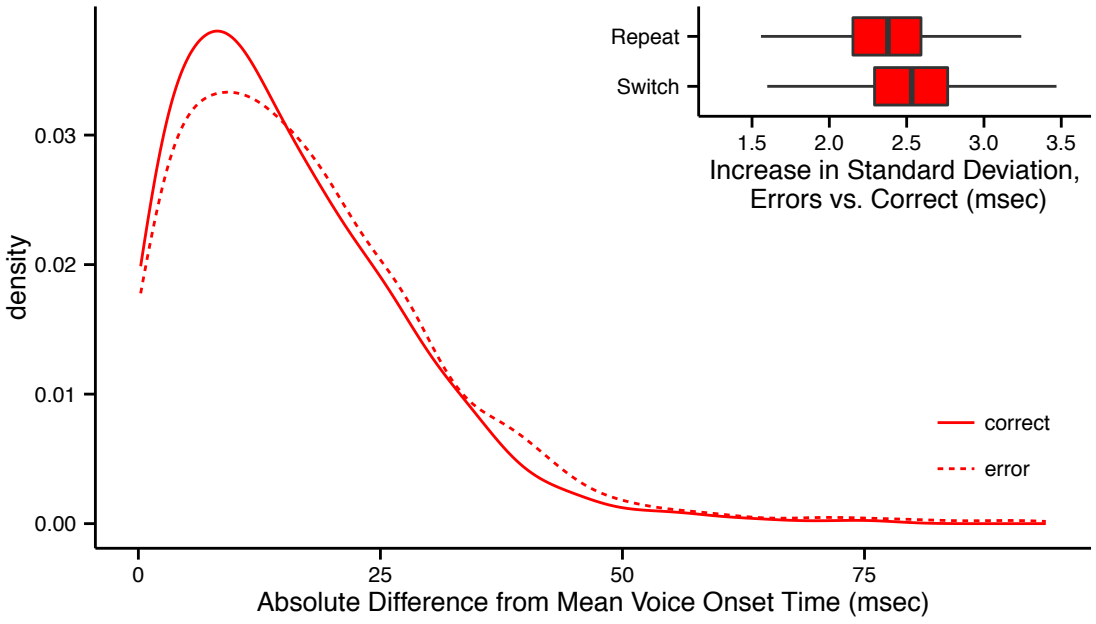
Across participants, no significant differences were observed across different types of twisters (95% confidence interval for difference, voiced consonants: [–0.03,0.7] milliseconds, $p > .05$; voiceless consonants: [–0.8, 1.1] milliseconds, $p > .30$); however, in the by-items analysis there was a tendency for variation to be lower in the repeat vs. switching condition (voiced 95% confidence interval: [–1.4, –0.5] milliseconds for voiced consonants, $p < .0001$; voiceless: [–1.5, –0.04] milliseconds for voiceless consonants, $p < .05$).

Table 4. Mean standard deviation of VOT (across participants) for errors in each tongue twister syllable ordering condition (95% confidence intervals for difference from correct productions in brackets).

	Voiced	Voiceless
Repeat	10.5 [2.2, 2.7]	18.8 [1.8, 3.0]
Switch	10.4 [2.6, 3.0]	18.9 [1.8, 3.2]



A.



B.

Figure 4. Increase in variance, errors vs. correct productions (A: voiced consonants; B: voiceless consonants). Smoothed density plots depict the absolute difference of each observation from the mean VOT for errors vs. a single random sample of correct productions produced by the same speaker on the same trials. Inset boxplots show the distribution of differences between the mean standard deviation of VOT (across participants) for errors and the mean standard deviation of 1,000 random samples of matched correct productions.

3.4.1 Power analysis of phonetic variability

To determine if our novel analysis method provides for a more accurate assessment of phonetic variability, we examined if datasets comparable in size to manually annotated studies would be able to detect the effect.

Excepting one large study examining 19,000 productions (Goldrick, et al., 2011; as noted in the introduction, requiring 3,000 person-hours for analysis), previous studies that manually annotated the acoustic properties of speech errors (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; McMillan & Corley, 2010) have analyzed data from ~8 participants, incorporating ~2,300 observations in total. To compare our results to these studies, for both voiced and voiceless consonants we randomly selected 10 subsets of 8 participants. We analyzed data from the first 28 trials (yielding 2,688 observations total).

We then repeated the analyses above on these subsets (using 100 random samples to estimate the chance distribution). As in the overall analysis, cases where the number of errors exceeded the number of matched correct productions were excluded. Given that only a small fraction of trials were used, this resulted in the exclusion of a greater proportion of errors. We restricted our analysis to random subsets where less than 12% of errors were excluded. Because variance was being analyzed, we also restricted our analysis to those participants that produced at least 2 errors in the first 28 trials.

There was a clear reduction in power. For voiced consonants, only 6/10 random subsets could recover the significant difference in variability observed in our large set of data. For voiceless consonants, only 5/10 random subsets recovered the difference in variability. Analyzing a large dataset provides us with the power needed to examine subtle patterns of variation in the phonetics of speech.

4. Discussion

Previous research on phonetic variation in speech production has been hamstrung by the reliance on subjective, labor-intensive manual annotation. Our approach provides a fully replicable method for rapidly analyzing acoustic data. Applying this to speech errors, our results provide further evidence against the traditional claim that speech errors are categorical substitutions of one sound for another (Dell, 1986; Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979). Our ability to analyze large amounts of data provided new insights into the how speech errors differ from categorical substitutions. Automatic analysis allowed us to collect sufficient data to compare types of twisters that had been utilized in different studies, but never directly contrasted. This revealed that some errors are closer to categorical substitutions than others. Errors exhibited smaller deviations from correct productions when twisters involved a switching vs. repeating alternation pattern. Second, analysis of a very large set of productions allowed us to detect that errors exhibit higher variance than correct production—an observation that would exceed the power of typical studies performed using manual annotation.

How can we understand these patterns within current theories of speech production? We attribute the difference in tongue twister orders to distinctions in the degree of gradience in planning vs. articulatory processes. Because the switching order requires alternation between distinct speech plans, increasing demands on planning processes (Rosenbaum, Weber, Hazelett, & Hindorff, 1986), it yields a greater number of errors within speech planning processes. Consistent with this, we find that errors increase at the point of switching between plans (i.e., at the first and third elements in the twister sequence). Assuming that gradience within planning processes is constrained—preferring relatively discrete representations over arbitrary blends

(Dell, 1986; Goldrick & Chu, 2014; Plaut & Shallice, 1993; Smolensky et al., 2014)—we would expect these errors to closely approximate categorical substitutions of one target for another. In contrast, errors arising in the continuous coordination of gestures during articulation should be less biased towards purely categorical substitutions. A *bin→pin* error should therefore be less /b/-like—more like a categorical substitution of a /p/—in the switching vs. repeating order.

The increased variability of errors may be attributed to the reduction of resources available for planning and articulation. This reduction in processing resources would result in mis-selection of incorrect speech plans as well as difficulty in implementing the appropriate articulations. These articulatory difficulties should be reflected in less precise, more variable phonetic properties for errors vs. correct productions.

In sum, we suggest that these findings support an integrated account of phonetic deviations. Such effects arise within gradient planning representations (Goldrick & Blumstein, 2006; Goldrick & Chu, 2014; Pouplier, 2007; Smolensky et al., 2014), sensitive to distinctions between sound categories, as well as within articulatory processes that execute such plans (Goldstein et al., 2007).

There are several clear avenues for extending our computational analysis methods. While we have focused on one specific aspect of the acoustic signal (VOT), this method can be applied to any acoustic dimension for which there are reliable analysis algorithms. Simultaneously examining more than one acoustic feature would allow us to take into account the multi-dimensional nature of speech sound contrasts (see Toscano & Murray, 2010, for a recent review and discussion). Extensions to the dynamic alignment process would allow for analysis of elicitation tasks that place fewer restrictions on the speaker, allowing us to move towards detailed automated analysis of spontaneous speech.

These findings illustrate the novel insights into language production that can be facilitated by the automatic analysis of large samples of speech. Such analyses could enhance many aspects of language production research. Returning to two examples discussed in the introduction, our understanding of how word duration reduces across repetitions (Kahn & Arnold, 2012, Lam & Watson, 2010) could be enhanced by systematic examination of a wide array of delays between mentions of a word. Understanding the mechanisms underlying accent variation following language switching (Balukas & Koops, 2015; Goldrick et al., 2014; Olson, 2013) could be enhanced by examining a wider array of bilingual speakers (who vary in second language proficiency, practice in switching, etc.). More generally, revealing the extent to which continuous variation in different cognitive processes modulates articulation and acoustics will help inform the development of theories incorporating gradient cognitive representations. Currently, such theories have high degrees of freedom. Gradient representations can specify a wide array of distinct representational states, and the relationship of such states to detailed aspects of speech has not been clearly specified (Pouplier & Goldstein, 2010, 2014). A richer empirical base will provide more constraints on such proposals.

Acknowledgments

Supported by National Science Foundation Grant BCS0846147 and National Institutes of Health Grant HD077140. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do no necessarily reflect the views of the NSF or the NIH. Thanks to the Northwestern SoundLab and Jennifer Culbertson for helpful discussion and comments.

References

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism, 19,* 423-443.

Boyce, S., Fell, H., MacAuslan, J., & Wilde, L. (2010). A platform for automated acoustic analysis for assistive technology. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies* (pp. 37-43). Association for Computational Linguistics.

Brugnara, F, Falavigna, D., Omologo, M. (1993). Automatic segmentation and labeling of speech based on hidden Markov models. *Speech Communication, 12*, 357–370.

Cho T., & Ladefoged, P. (1999) Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics, 27*, 207–229.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006) Online passive-aggressive algorithms. *Journal of Machine Learning Research, 7*, 551−585.

Croot, K., Au, C., & Harper, A. (2010). Prosodic structure and tongue twister errors. In C. Fougeron, B. Kuhnert, M. D'Imperio & N. Vallee (Eds.) *Laboratory phonology 10: Variation, phonetic detail and phonological representation* (pp. 433-461). Berlin: Mouton de Gruyter.

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

Dell, G. S., & O'Seaghdha, P. G. (1994). Inhibition in interactive activation models of linguistic selection and sequencing. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 409–453). San Diego: 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.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language, 66*, 789-806.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus [CD-ROM]. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N, 93, 27403.

Goldrick, M., Baker, H. R., Murphy, A., & Baese-Berk, M. (2011). Interaction and representational integration: Evidence from speech errors. *Cognition, 121*, 58–72.

Goldrick, M., Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649–683.

Goldrick, M. & Chu, K. (2014). Gradient co-activation and speech error articulation: Comment on Pouplier and Goldstein (2010). *Language, Cognition, and Neuroscience, 29*, 452–458.

Goldrick, M., Runnqvist, E., & Costa, A. (2014). Language switching makes pronunciation less nativelike. *Psychological Science, 25*, 1031-1036.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*, 386–412.

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

Hansen, J., Gray, S., & Kim, W. (2010). Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification. *Speech Communication, 52*, 777–789.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al. (1975). *Journal of Memory and Language, 52*, 58–70.

Heisler, L., Goffman, L., & Younger, B. (2010). Lexical and articulatory interactions in children's language production. *Developmental Science, 13*, 722-730.

Hosom, J. P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication, 51*, 352–368.

Kahn, J. M., Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language, 67*, 311-325.

Keshet, J., Shalev-Shwartz, S., Singer, Y., Chazan, D. (2007) Large margin algorithm for speech-to-phoneme and music-to-score alignment. *IEEE Transactions on Audio, Speech, and Language Processing, 15,* 2373-2382.

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and renanalysis. *Language, 89*, 30-65.

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory and Cognition, 38*, 1137-1146.

Lisker, L., Abramson, A. S. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

MacDonald, P. (2012) *Mixdist: Finite mixture distribution models*. (R package version 0.5-4; http://cran.r-project.org/web/packages/mixdist/).

McAllester, D., Hazan, T., & Keshet, J. (2010). Direct loss minimization for structured prediction. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., & Culotta, A. (Eds.) *Advances in Neural Information Processing Systems 23* (pp 1594-1602). Neural Information Processing Systems Foundation.

McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition, 117*, 243–260.

McMillan, C. T., Corley, M., & Lickley, R. J. (2009). Articulatory evidence for feedback and competition in speech production. *Language and Cognitive Processes, 24*, 44–66.

Olson, D. J. (2013). Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics, 41*, 407-420.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*, 377-500.

Pouplier, M. (2003). *Units of phonological encoding: Empirical evidence*. Unpublished doctoral dissertation, Yale University, New Haven, CT.

Pouplier, M. (2007) Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech, 50*, 311–341.

Pouplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics, 36*, 114–140.

Pouplier, M. & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. *Language and Cognitive Processes, 25*, 616–649.

Pouplier, M. & Goldstein, L. (2014). The relationship between planning and execution is more than duration: Response to Goldrick & Chu. *Language, Cognition, and Neuroscience, 29*, 1097–1099.

Prathosh A P, Ramakrishnan A G, Ananthapadmanabha T. V. (2014) Estimation of voice-onset time in continuous speech using temporal measures. Journal of the Acoustical Society of America 136:EL122-128.

Rosenbaum D A, Weber R J, Hazelett W M, Hindorff V (1986) The parameter remapping effect in human performance: Evidence from tongue twisters and finger fumblers. Journal of Memory and Language 25: 710–725.

Ryant, N., Yuan, J., Liberman, M. (2013). Automating phonetic measurement: The case of voice onset time. *Proceedings of Meetings on Acoustics, 19*, 060277.

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

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.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science, 38*, 1102–1138.

Sonderegger, M., & Keshet, J. (2012.) Automatic measurement of voice onset time using discriminative structured prediction. *Journal of the Acoustical Society of America, 132,* 3965-3979.

Stouten, V., & van Hamme, H. (2009). Automatic voice onset time estimation from reassignment spectra. *Speech Communication, 51*, 1194-1205.

Toscano, J., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34*, 434-464.

Wesenick, M.-B. and Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proc. 4th International Conference on Spoken Language Processing (ICSLP)*.

Wilshire, C. E. (1999). The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech, 42*, 57–82.

Yuan, J, & Liberman, M. (2014) F_0 declination in Mandarin broadcast news speech. *Speech Communication, 65*, 67-74.

Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences USA, 107*, 592-597.

Appendix

Table A1. Quadruplets of pairs of syllables that form the basis for tongue twisters. Each pair in a quadruplet generates 4 tongue twisters (see text for details).

Pair 1		Pair 2	
Word 1	Word 2	Word	Nonword
punk	bunk	pulp	bulp
punch	bunch	pulse	bulse
pox	box	posh	bosh
peat	beat	peal	beal
bowl	pole	bone	pone
beak	peek	bean	pean
bet	pet	bell	pell
boast	post	bolt	polt
tab	dab	tat	dat
tense	dense	tenth	denth
torque	dork	torn	dorn
tote	dote	toad	dode
dour	tower	douse	touse
dense	tense	dealt	telt
dart	tart	darn	tarn
dune	tune	dupe	toop
cod	god	cop	gop
cuff	guff	cud	gud
cape	gape	cake	gake
code	goad	comb	gome
gap	cap	gas	cass
goon	coon	goose	coose
guilt	kilt	gift	kift
gain	cane	gate	kate

Lexicality analyses

1. Error rates

As shown in Table S1, participants' accuracy rate was roughly 90% across conditions. (Confidence intervals in Table S1 and throughout were estimated using a bootstrap with 1,000 samples.) We unexpectedly failed to observe a reliable lexical bias effect, where errors favor word over nonword outcomes. Accuracy was not substantially lower on nonword targets (where a voicing error would create the word outcome favored by this bias), nor was it substantially higher on the corresponding word target (where a voicing error would produce a disfavored nonword outcome). It is unclear why this was the case. Previous work has suggested the lexical bias effect weakens with faster response rates (Dell, 1986); one possibility is that the tongue twister production rate was too rapid for lexical effects to observed.

Table S1. Mean (across participants) proportion correct on elements of quadruplets (estimated 95% confidence interval in brackets).

Condition	Word 1	Word 2	Word	Nonword
Repeat	89.8% [87.7%,91.8%]	91.3% [89.2%,93.0%]	90.2% [88.0%,92.3%]	91.2% [89.2%,93.0%]
Switch	88.0% [84.5%,91.0%]	91.2% [88.7%,93.5%]	88.2% [85.2%,91.0%]	88.9% [86.1%, 91.5%]

2. VOT analysis

Previous work (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; McMillan, Corley, & Lickley, 2009) has suggested that whether or not an error results in a word vs. a nonword can influence the degree to which an error deviates from a corresponding correct production. Studies with nonword targets (Goldrick & Blumstein, 2006; McMillan et al., 2009) have shown that errors resulting in nonwords (*keff*→*geff*) show larger deviations from correct outcomes than errors resulting in words (*keese*→*geese*). Consistent with this result, one study examining word

targets (Frisch & Wright, 2002) provided some evidence that errors resulting in nonwords (*suck→zuck*) are more likely to result in tokens with atypical phonetic properties relative to errors resulting in words (*sue→zoo*).

Parallel to this latter study, by contrasting syllables across pairs within a quadruplet, we can examine how the lexicality of the outcome influences errors on word targets. Monte Carlo analyses structure following those reported in the main text estimated the 95% confidence intervals for correct productions matched to two types of errors. Tables S2-S3 compare errors on word targets that result in nonwords (*bolt→polt*) to matched errors that result in words (*boast→post*).

Table S2. Mean VOT (across participants) of errors in conditions contrasting in outcome lexicality (95% confidence intervals for difference from correct productions in brackets).

Condition	Word Target → Nonword Error Outcome	Matched Word Target → Word Error Outcome
Voiced	28.0 [4.9, 5.7]	29.3 [6.5, 7.2]
Voiceless	52.4 [–16.7, –14.9]	52.9 [–15.9, –13.9]

Table S3. Mean VOT (across quadruplets) of errors in conditions contrasting in outcome lexicality (95% confidence intervals for difference from correct productions in brackets).

Condition	Word Target → Nonword Error Outcome	Matched Word Target → Word Error Outcome
Voiced	27.8 [4.6, 5.4]	29.0 [6.3, 7.0]
Voiceless	50.5 [–16.5, –15.0]	51.0 [–16.0, –14.5]

We failed to find a consistent effect across voiced and voiceless consonants. In the by-participants analysis (Table S2), voiced outcomes resulting in nonwords showed smaller deviations from correct productions than errors resulting in words (95% confidence interval for

difference across conditions [–2.1, –1.0] milliseconds, $p < .0001$) whereas there was a non-significant difference for voiceless consonants (95% CI [–2.4, –0.5] milliseconds, $p > .05$; recall for voiceless consonants that a negative difference indicates a stronger deviation from correct productions). Similar results were found in the by-quadruplets analysis (Table S3; 95% confidence interval for difference across conditions, voiced consonants [–2.2, –1.1] milliseconds, $p < .0001$; 95% CI for voiceless consonants [–1.6, 0.6] milliseconds, $p > .05$).

Our design also allows us to examine how the lexicality of the target influences errors resulting in word outcomes (unexamined in previous work). Tables S4-S5 compare errors on nonword targets that result in words (*polt*→*bolt*) to matched errors on word targets (*post*→*boast*). For both voiced and voiceless consonants, we find that deviations from correct productions are larger for nonword→word relative to word→word errors.

Table S4. Mean VOT (across participants) of errors in conditions contrasting in target lexicality (95% confidence intervals for difference from correct productions in brackets).

Condition	Nonword Target → Word Error Outcome	Matched Word Target → Word Error Outcome
Voiced	29.4 [6.2, 7.1]	28.1 [5.2, 6.2]
Voiceless	52.7 [–11.0, –9.0]	54.8 [–4.2, –1.3]

Table S5. Mean VOT (across quadruplets) of errors in conditions contrasting in target lexicality (95% confidence intervals for difference from correct productions in brackets).

Condition	Nonword Target → Word Error Outcome	Matched Word Target → Word Error Outcome
Voiced	28.2 [4.9, 5.8]	28.8 [4.0, 5.1]
Voiceless	50.2 [–15.1, –13.5]	51.4 [–12.3, –10.5]

In the by-participants analysis (Table S4), errors on nonword targets showed larger deviations from correct productions than errors on word targets (95% confidence interval for difference across conditions, voiced consonants: [0.3, 1.5] milliseconds, $p < .0001$; voiceless consonants: [–4.2, –1.3] milliseconds, $p < .0001$). Similar results were found in the by-quadruplets analysis (Table S5; 95% CI for difference across conditions, voiced consonants: [0.02, 1.5] milliseconds, $p < .05$; voiceless consonants: [–4.1, –1.7] milliseconds, $p < .0001$).

This latter result provides some additional qualified support for the role of planning processes in speech error articulation. Lexicality reflects a property of sound sequences that is stored in long term memory—their association to lexical items and meanings. This property of memory modulates the degree to which errors deviate from correct productions. However, it is unclear why nonword targets would exert a stronger influence on error articulation than comparable word targets. Given our failure to observe an influence of lexicality on the probability of error outcomes, we can speculate that some aspect of the experimental context may have served to strengthen the activation of nonwords (eliminating the advantage for word outcomes and strengthening their influence on error articulation). Examination of this possibility requires further manipulation of the experimental context in which tongue twisters are elicited.

References

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

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.

Goldrick, M., Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649–683.

McMillan, C. T., Corley, M., & Lickley, R. J. (2009). Articulatory evidence for feedback and competition in speech production. *Language and Cognitive Processes, 24*, 44–66.

Wilshire, C. E. (1999). The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech, 42*, 57–82.